Informfully Recommenders – A Reproducibility Framework for Diversity-aware Intra-session Recommendations

Anonymous Author(s)

Abstract

In recent years, norm-aware recommender systems have gained increased attention, especially for diversity optimization. The recommender systems community has well-established experimentation pipelines that support reproducible evaluations by facilitating models' benchmarking and comparisons against state-of-the-art methods. However, to the best of our knowledge, there is currently no reproducibility framework that supports thorough norm-driven experimentation at the four stages of the recommender pipeline: pre-processing, in-processing, post-processing, and evaluation stages. To address this gap, we present Informfully Recommenders, a first step towards a normative reproducibility framework that focuses on diversity-aware design built on Cornac. Our extension provides an end-to-end solution for implementing and experimenting with normative and general-purpose diverse recommender systems that cover 1) dataset pre-processing, 2) diversity-optimized models, 3) dedicated intra-session item re-ranking, and 4) an extensive set of diversity metrics together with item visualization for offline and online evaluation. We demonstrate the capabilities of our diversity-aware extension—and in particular of our diversity-driven recommendation models—by providing an extensive offline experiment in the news domain.

CCS Concepts

• **Information systems → Recommender systems**; **Information retrieval diversity**; • **Human-centered computing → Open source software**.

ACM Reference Format:
Anonymous Author(s). 2025. Informfully Recommenders – A Reproducibility Framework for Diversity-aware Intra-session Recommendations. In *Proceedings of The 19th ACM Recommender Systems Conference (RecSys '25).* ACM, New York, NY, USA, 10 pages. https://doi.org/TBA

1 Introduction

Recommender systems (RS) can be powerful tools to help users find their way in the vast information space found online, shape public discussion, and serve as a foundation for public cohesion [8, 25, 28]. In the news domain, for example, they fulfill an important "democratic role" [27] for political opinion formation. Here, content curation by RS has a direct impact on the human right of freedom of expression through informational self-determination [48]. Depending on the target domain, RS therefore face a unique set of challenges that require balancing societal norms and values on the one hand [28] as well as technical performance and economic goals of platform owners on the other hand [17, 46]. In the context of this paper, we refer to such RS instances as *normative* RS (NRS). We follow the definition of Vrijenhoek et al. [60] where normativity is understood as the practice of operationalizing societal norms and values as part of the RS pipeline.

Development and evaluation of NRS is challenging. When looking at a target objective for NRS such as diversity or serendipity, there is disagreement on the conceptual level on the precise notion [37, 58]. Additionally, online user studies for assessing and evaluating the algorithm's impact on users remain the exception in RS research [6]. These studies, however, are vital for assessing the performance of NRS, because with the predominant focus on offline evaluation, it is unclear how algorithms impact/matter to users [31]. As a result, the understanding of norm-aware models and/or re-rankers remains limited [52]

One reason for this lack of proper assessment is the requirement to have sufficiently rich datasets with complementary information on items, participants' backgrounds [26, 29], normative models/re-rankers [11], diversity metrics [58], and visualizations [7, 22] for the evaluation with users. Despite recent efforts to promote more normative and beyond accuracy perspectives (e.g., see RecSys 2024 Challenge [23, 32, 33]), there have not yet been any dedicated end-to-end pipelines proposed for the systematic evaluation of NRS.

To address these shortcomings, we present Informfully Recommenders, a first reproducibility framework for norm-aware approaches, such as diversity.[1] It is designed as an extension to the well-established Cornac framework [47, 53] and provides an end-to-end solution for implementing and experimenting with NRS. Our framework's contributions include: 1) six out-of-the-box dataset augmentation functions for adding norm-relevant dimensions to items in (supporting multiple languages), 2) three random walks and two lightweight diversity models for generating norm-aware recommendations, 3) four intra-session diversity re-rankers together with a user simulator for diversity-optimized re-ranking; 4) ten traditional and normative diversity metrics for assessing recommendations, and 5) compatibility with the Informfully research framework [22] for item visualization that is ready to be deployed as a back end to support online user studies.

This enables Informfully Recommenders to assist the systematic integration, evaluation, and assessment of societal values and norm-awareness into recommender systems. We show the applicability of our extension in the context of news recommendations—a domain closely linked to normative societal values in general [27, 59, 60] and diversity in particular [23, 57].

RecSys '25, Sep. 22–26, 2025, Prague, Czech Republic

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/TBA

[1] Informfully Recommenders: https://github.com/Informfully/Recommenders
Experiment configuration files: https://github.com/Informfully/Experiments
Online code documentation: https://informfully.readthedocs.io

Table 1: Overview of open-source reproducibility framework. The comparison looks at supported datasets, models, re-rankers, and metrics, as well as augmentation, simulation, and visualization capabilities.

Framework	Modes	Models	Re-rankers	Metrics	Data Augmentation	User Simulator	Item Visualization
ClayRS [38]	OFF	TRA	N/A	ACC			
Cornac [47, 53, 54] + A/B [41]	ON, OFF	TRA	N/A	ACC, BEY			✓
daisyRec 2.0 [50]	OFF	TRA	N/A	ACC, BEY			
Elliot [4]	OFF	TRA	N/A	ACC, BEY			
FuxiCTR [68, 69]	OFF	TRA	N/A	ACC			
LensKit [19]	OFF	TRA	N/A	ACC			
Microsoft Recommenders [5]	OFF	TRA	N/A	ACC, BEY			
RecBole [67]	OFF	TRA	N/A	ACC, BEY	✓		
ReChorus 2.0 [35]	OFF	TRA	STA	ACC			
RecList [14]	OFF	TRA	N/A	N/A		✓	
RecPack [40]	OFF	TRA	STA	ACC, BEY	✓		
Informfully Recommenders	ON, OFF	NOR, TRA	STA, DYN	ACC, BEY	✓	✓	✓

In this paper, we first compare existing tools in Section 2 and identify their shortcoming for supporting norm-aware user experiment. We then outline our framework extension in Section 3, provide a sample experiment configuration in Section 4, discuss the results in Section 5 and a conclude with an outlook in Section 7.

2 Related Work

In this section, we compare the available RS reproducibility framework offerings to show their respective shortcomings for assessing NRS. Table 1 presents a comparison of open-source reproducibility frameworks that can be used to tackle news recommendations and diversity.[2] We compare the capabilities of the frameworks on the following dimensions:

- **Modes:** Shows if frameworks support **on**line user experiment (ON) or if they are focusing on **off**line benchmarking (OFF).
- **Models:** Lists available model types. Options include **nor**mative (NOR) and **tra**ditional models (e.g, accuracy, TRA).
- **Re-rankers:** We differentiate between **sta**tic re-ranking of candidate lists after the model stage (STA)[3] and **dyn**amic intra-session re-ranking that takes user interactions into account and can iterate multiple times (DYN). Frameworks without dedicated re-ranking steps read "N/A."
- **Metrics:** Shows if generic **acc**uracy (ACC) or norm-relevant **bey**ond accuracy metrics (BEY) are included. Reads "N/A" if no metrics are available.
- **Augmentation, Visualization, Simulator:** Furthermore, our comparison covers three additional dimensions critical for NRS: 1) data augmentation (to add required normative attributes for models, re-rankers, or metrics), 2) user simulator (for offline benchmarking of intra-session re-ranking), and 3) item visualization (for conducting user studies). We mark an entry with ✓ if they include data augmentation, a user simulator, or item visualization.

[2]This list is based on the ACM RecSys repository of evaluation framework recommendations: https://github.com/ACMRecSys/recsys-evaluation-frameworks

[3]The re-ranker must be part of a modularized pipeline. Re-rankers tied to/part of a model cannot be reused across different algorithms and are listed as "N/A".

Looking at Table 1, we see that the support of online experimentation is limited to one framework. We do not find any support for norm-aware models. The same goes for re-ranking, where all but one framework do not even include a dedicated stage for this process. The situation is better for metrics, where six frameworks include beyond accuracy assessment that include diversity.

Data Augmentation: Experimentation with NRS is heavily dependent on rich contextual information, as models, re-rankers, or metrics require, e.g., information on political actors or item embeddings to work. This data, however, is rarely present in datasets. Table 1 shows that only two framework offers built-in functionality for data augmentation to add such norm-relevant information.

User Simulators: Intra-session effects can have a significant impact on user engagement [39]. Using sequence-aware information on intra-session behavior offers a rich source of information to personalize recommendations [51]. However, leveraging intra-session data for news recommendations is among the least popular domains in sequence-aware RS research [45].

While there exists previous work on user simulation for recommender system, they are either conceptual or theoretical in nature without implementation [20], stand-alone implementations that are not part of any testing framework [13, 30, 66], tied to a specific domain [1], or a combination of simulators and re-rankers [65]. We see this also reflected in Table 1, where only a single framework offers the capability of testing with simulated user interactions.

Item Visualization: Only the extended Cornac A/B supports item visualization for user experiments. While we use the same underlying Cornac back end, our extension utilized Informfully [22], a more generic approach that is not tied to a specific framework. This allows our framework to act as a back end for supporting online user studies by taking the recommendation lists and forwarding them to an application where people can interact with them.

To the best of our knowledge, there is no open-source reproducibility framework that offers a unified, end-to-end solution across the four main RS stages with dataset operations, model selection, (intra-session) re-ranking with user simulation, and metrics assessment together with item visualization.

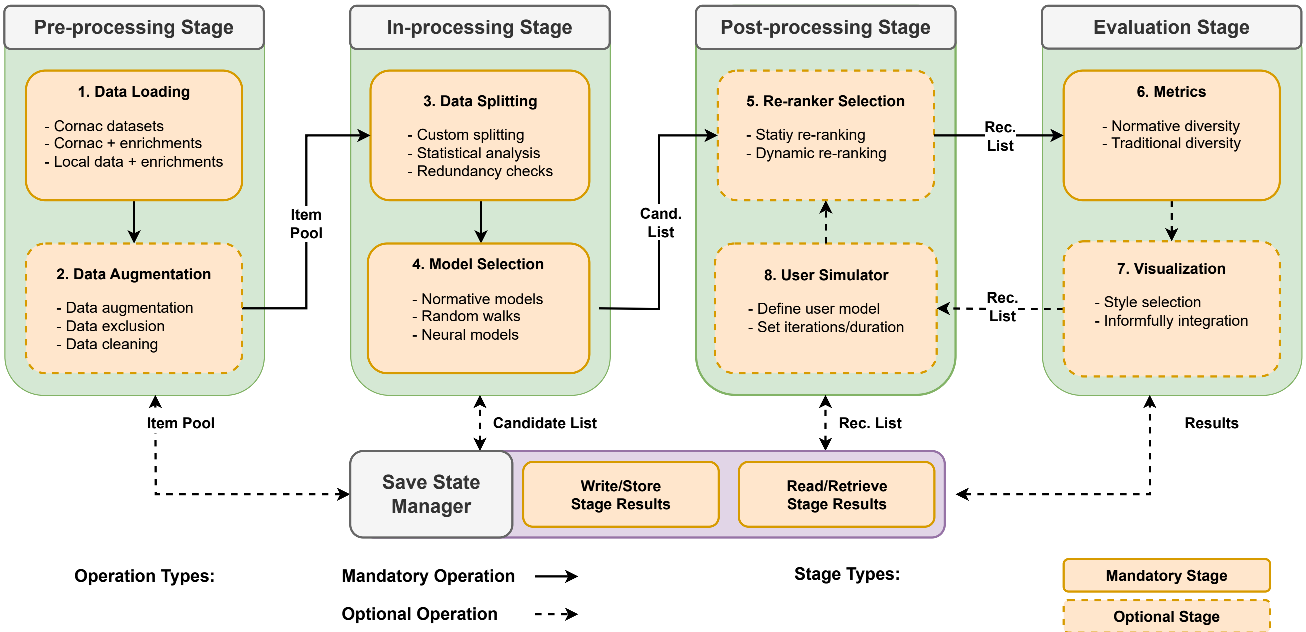

Figure 1: Informfully Recommenders extension of the existing Cornac pipeline by implementing a diversity-aware four-stage RS pipeline with eight customizable steps. It includes a Save State Manager for saving and loading results at each stage. Information gets passed across stages with specific files (i.e., item pool, candidate lists, and recommendation lists).

3 Informfully Recommenders

Informfully Recommenders is an extension of the Cornac framework for multimodal RS [47, 53, 54]. Figure 1 provides an overview of the updated reproducibility framework. Following the modularized stages of Table 1, our norm-aware diversity extension of Cornac splits the RS pipeline into the four stages of pre-processing (dataset operations), in-processing (model operations), post-processing (re-ranking operations), and evaluation (metrics assessment and visualization). Each stage is further sub-divided into two steps, allowing researchers to customize the intended behavior of the RS.

The entire pipeline (cf. Figure 1) is modularized and each stage represents a self-contained unit. Communication between stages is done exclusively via exchanging item files (i.e., Item Pool, Candidate List, and Recommendation List—solid arrows in Figure). For convenience, we provide a Save State Manager for storing and retrieving these item files. This manager allows the pipeline to be initialized at any stage by reusing existing intermediate results, speeding up the development process (e.g., when testing multiple re-ranking approaches, one candidate list is sufficient as it can be (re-)loaded for subsequent re-ranking rounds, skipping the pre- and in-processing stages of the pipeline).

Our extended framework presents a complete end-to-end pipeline for RS. It allows for reproducibility across the in-processing [61], post-processing [44], and the evaluation stage [21]. These capabilities allow Informfully Recommenders to be used for both general-purpose and diversity-driven offline benchmarking/development purposes, as well as being deployed as a back end for conducting user studies; it has a successful track record of powering online user studies (for more details, please see [23–26]). Furthermore, by making the newly introduced stages and steps optional, we ensure full backwards compatibility with existing Cornac experiments.

3.1 Pre-processing Stage

The purpose of the pre-processing stage is to prepare the user-item interactions and to define the item pool. The extension to the pre-processing stage includes two main additions: 1) customizable data loading options and 2) data augmentation functions.

Data Loading: Cornac uses a user-item rating matrix to load and process data. The main limitation here is that it does not contain any date information on when an interaction took place.[4] We now allow users to load separate history files during the in-processing model stage via the *userHistory* parameter that can contain custom attributes. And looking at the final recommendation output, Cornac only considers items that appear both in the training *and* the test sets for recommendation purposes. However, this can be too broad or too narrow depending on the specific use case. We therefore extended the base recommendation model and re-ranker with an optional *articlePool* parameter. This pool can be used to either extend or reduce the items for which we calculate a prediction in the candidate list. Both *userHistory* and *articlePool* are optional parameters to make the extension fully backwards compatible with existing Cornac experiments.

[4] This time component, however, can be crucial information, as it allows to, e.g., discount older interactions and/or track the status of impression lists. Depending on the domain, this presents valuable information that the models could leverage.

Data Augmentation: The data augmentation extension is an optional step comprising several *text* enrichment functions, such as sentiment analysis, named entity recognition (NER), the identification of political actors and parties, assessment of text complexity, identification of event clusters, and categorization of item types. The data augmentation pipeline supports texts in multiple languages, including English, German, Danish, and Portuguese.

Sentiment Analysis: We offer sentiment analysis for texts using roBERTa.[5] The sentiment ranges from negative (-1.0) to positive (1.0). We take this score and group articles into four baskets, expressing an opinion that is either "negative," "somewhat negative," "somewhat positive," or "positive."[6]

Named Entity Recognition: Named entities of various types (e.g., people, locations, organizations, events, among others) are extracted using the spaCy library.[7]

Political Actors: The political augmentation identifies politicians and parties using a combination of spaCy for NER and Wikidata[8] for further augmenting the named entities with politics-centric information. By default, political annotations are categorized using a three-way split and they are assigned to either the "Majority Party," "Minority Party," or "Others/Foreign Parties' bins.

Text Complexity: The framework assesses the complexity of a text using the Textstat library.[9]

Event Clusters: We calculate event clusters to allow grouping texts, such as news articles, by story based on text similarity and named entities within categories using NetworkX.[10]

Text Categories: This augmentation feature allows for automatically assigning a category to a text using BART.[11]

Helper Function: We include auxiliary functions to clean and validate the original as well as the augmented data files. Options include filtering invalid articles (i.e., items with empty attributes) and removing users or items with no recorded interactions/history. After validating all the articles, the script can prepare the user-item rating matrix required by Cornac.

We provide ready-to-go augmentation pipelines and sample code for the Ekstra Bladet News Recommendation Dataset (EB-NeRD, Danish) [32], the Microsoft News Dataset (MIND, English) [64], and the German News Collection on Migration (NeMig, German) [29] to showcase how the added augmentation steps perform across different languages and datasets.[12]

3.2 In-processing Stage

The purpose of the in-processing stage is to generate a candidate list of items from the item pool. The extended framework contains five new data splitting methods and three new families of algorithms that allow for experimenting with both diversity-driven recommendations and news recommendations.

Data Splitting: We introduce five additional data splitting methods: attribute-based sorting, diversity-based subset construction, attribute-based stratified splitting, diversity-based stratified splitting, and a clustering-based approach. The main motivation behind these splitting methods is not primarily the improvement of target metrics. Instead, the goal is to see how the model's performance is affected by changes in the underlying dataset.

Attribute-based Sorting: Allows sorting by item or user attributes before splitting, e.g., by article sentiment, to see how the resulting recommendation changes if the training set mainly consists of articles of a particular sentiment.

Diversity-based Subset Construction: Construction of an item subset for training and testing with a purposefully skewed diversity across a target dimension to ascertain how this imbalance impacts recommendations.

Attribute-based Stratified Splitting: Stratified splitting that allows for the generation of train and test sets with balanced item attributes (e.g., equal distribution of political parties).

Diversity-based Stratified Splitting: Measures users' diversity (e.g., in terms of political viewpoints) and controls their distribution across training and test sets.

Clustering-based Stratified Splitting: Using K-means and PCA clustering approaches to control the homogeneity of training and test sets.

Model Selection: Our framework extends Cornac with three families of algorithms: 1) five neural models from the literature, 2) our norm-aware filtering algorithms (for both on- and offline use), and 3) three random-walk-based approaches.

Neural Models: Our framework extension includes five **non-normative** neural models from the literature. These neural models are meant to be used as baselines and were selected due to their satisfactory performance in past news recommender systems challenges [64]. We included the Efficient Neural Matrix Factorization (ENMF) [12], Long- and Short-Term User Representation (LSTUR) [2], variational autoencoders (VAE) [36], Neural News Recommendation with Personalized Attention (NPA) [62], and with Multi-Head Self-Attention (NRMS) [63].

Filtering Algorithms: The filtering algorithms present algorithms that incorporates social norms and values by using a normative target distribution (NTD). A NTD is a list of item attributes, relevant attribute values, and the overall occurrence count of these values.[13] We include two lightweight, **normative** filtering algorithms from the literature, namely participative Political Diversity (PLD) [25] and deliberative Exposure Diversity (EPD) [24] as outlined by Helberger [27]. PLD implements a model *participatory* understanding of democracy. It generates recommendations by focusing on articles that share the main political viewpoints in society, with the goal of creating a common background knowledge for participating in political discussions. EPD implements a model of the *deliberative* understanding of democracy. Its

[5]The underlying model can be exchanged. Our sample implementation uses XLM-roBERTa: https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[6]The number of baskets used here serves as an example. This can be customized to fit the specific experimental requirement/setting.
[7]NER with spaCy: https://spacy.io/usage/linguistic-features#named-entities
[8]Wikidata website: https://www.wikidata.org
[9]Textstat library: https://pypi.org/project/textstat
[10]NetworkX documentation: https://python-louvain.readthedocs.io
[11]BART model: https://huggingface.co/facebook/bart-large-mnli
[12]The code is available: https://github.com/Informfully/Experiments

[13]For example, in the case of news, the editors can define an NTD for political parties mentions (item attribute) that gives party A and B (attribute values) the same exposure (50 mentions of party A and 50 mentions of party B) to ensure balanced reporting reflecting existing journalistic principles or regulatory requirements.

goal is to create inclusive recommendations and give exposure to all political positions in order to provide a diverse selection of perspectives for consensus seeking.

Random Walks: Our extension includes the **non-normative** random walk algorithm with popularity discount RP^3_β [15] and Random Walks with Erasure (RWE-D) [43], as they have shown to have excellent performance in the item diversification problem. Finally, we include Diversity-Driven Random Walk (D-RDW), a novel **normative** RS that capitalizes on the diversification capabilities of the traditional random walk algorithms and combines it with NTD.

3.3 Post-processing Stage

The goal of the post-processing stage is to offer a lightweight re-ranking option for recommendations to accommodate metric optimization and/or business logic. This is a new step in the Cornac pipeline. To make everything backward-compatible, we set default parameters for the new stages to run old experiments that do not explicitly specify a re-ranking step. In addition to one-time re-rankers (i.e., *static* re-ranking), this stage also includes a user simulator to support iterative re-ranking (i.e., dynamic re-ranking).

Re-ranker: Informfully Recommenders presents a fully customizable approach that allows for static (one-time heuristics/filters) and dynamic re-ranking (accounting for intra-session user interactions). The re-ranking logic allows for intra-session adjustments of the recommendation list in a static and dynamic fashion.

Static Re-rankers: The static re-rankers are applied to the candidate list right after the model step to optimize the output for a target metric such as diversity [11]. To that end, the output of the models can be re-ranked using three customized approaches: 1) Greedy-KL (with optional windows) [49], 2) PM-2 [18], and 3) MMR [10].[14]

Dynamic Re-rankers: Alternatively, we implemented a dynamic intra-session re-ranking option that updates recommendations based on user interaction (using items from the candidate list). The default strategy implemented in this framework is the dynamic attribute penalization (DAP). DAP diversifies the recommendation list by penalizing items in the upcoming session that share attributes with clicked items. DAP can be combined with heuristics.[15]

User Simulator: Dynamic re-ranking requires an underlying user model that specifies how the item feed is being browsed. We provide a sample template that can be customized and extended. In the context of NRS, the two default behaviors included in the framework are: 1) Users are more likely to click on articles from a category that they have previously read, and 2) Items higher up in the recommendation list are more likely to be clicked (cf. [65]). Apart from the interactions, the user models allow researchers to specify the overall duration and number of loops (i.e., how many times recommendations are calculated and "consumed" by the agent).

[14] G-KL and PM-2 use the same NTD sampler as D-RDW. MMR creates an equal distribution across the target dimensions.

[15] E.g., implemented a default rule that removes any items that were already clicked during the session.

3.4 Evaluation Stage

The evaluation stage includes the final steps of the recommendation pipeline. The two main contributions of our extension are 1) the addition of beyond accuracy metrics to assess the recommendation quality in terms of item diversity and 2) item visualization to show the system output to users for conducting online experiments to gather feedback.

Metrics: We integrate five rank-aware divergence metrics for measuring *normative diversity*, called the RADio metrics [57], namely calibration, fragmentation, activation, representation, and alternative voices. The normative RADio metrics are based on democracy theory [28] and are specially tailored for assessing the normative dimension of RS [57]. These metrics consider various item features, such as topics, sentiment, named entities, and political parties, as well as additional contextual information, such as the user history, the pool of available items, and the relevance of items. Furthermore, we implemented traditional diversity metrics such as intra-list distance and expected intra-list distance [9], Gini coefficient [11], α-nDCG [16], and binomial diversity [55]. These traditional diversity metrics are considered domain-agnostic and applicable to a wide range of domains [11].

Visualization: The last step of the NRS pipeline consists of visualizing the recommendation list. Our framework has built-in support for the Informfully Platform [22], by transforming recommendations into the JSON recommender exchange format (JREX) for presenting them to users.[16] This is an optional step that can be disabled for offline testing and benchmarking.

4 Experiments

We demonstrate the capabilities of Informfully Recommenders by running diversity-focused experiments using neural models, filtering algorithms, and random walks on reference news datasets.[17]

Datasets: In our experiments, we used three well-known news datasets, namely EB-NeRD (small version) [32], MIND (small version) [64], and NeMig (German subset) [29] to compare and evaluate a diverse range of models across normative and traditional diversity metrics. Table 5 provides an overview of the datasets.

We limit the data cleaning steps to removing users from the test set that are not part of the training set and removing items that have empty/no text attributes (e.g., movie trailers). We applied data augmentation steps, as outlined in Section 3.1. We performed NER to identify political actors and parties in articles and assign them to one of three buckets: majority, minority, or others.[18] Furthermore, we perform sentiment analysis to classify each article and apply event clustering to identify articles that cover similar stories. This is all done using the newly added data augmentation functions.

[16] Online tutorial: https://informfully.readthedocs.io/en/latest/recommendations.html

[17] News recommendations become increasingly popular in the RS community (cf. RecSys Challenge 2024 [32, 33]). And using them as an example allows us demonstrate the new elements from each of the pipeline stage.

[18] We note here that MIND is a dataset from the US. Thus, the majority parties here are the *Democratic Party* and *Republican Party*. Other national parties are considered minority parties. The third bucket for others is limited to foreign parties. For EB-NeRD and NeMig, the majority bucket includes all governing parties, and the minority bucket contains all opposition parties. The remaining actors are in the third bucket.

Table 2: Overview for *EB-NeRD* of the diversity scores for the top 20 news recommendations and AUC scores for predicting user impressions. The values closest to the perfect score are highlighted in green, second closest in blue, and third closest in red.

Model	Re-ranking	Activ.	Cat. Calib.	Comp. Calib.	Frag.	Alt. Voices	Repr.	Cat. Gini	Sent. Gini	Party Gini	Cat. ILD	Sent. ILD	Party ILD	Train. Cost	Rec. Cost	AUC
Perfect (NTD)	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.133	0.250	1.000	0.779	0.789	N/A	N/A	1.000
LSTUR [3]	N/A	0.190	0.433	0.257	0.610	0.063	0.372	0.832	0.645	0.874	0.740	0.552	0.343	1112.29	69.84	0.564
NPA [62]	N/A	0.202	0.468	0.266	0.619	0.061	0.365	0.826	0.605	0.877	0.753	0.580	0.337	1656.96	51.25	0.554
NRMS [63]	N/A	0.204	0.509	0.285	0.632	0.060	0.362	0.791	0.544	0.880	0.794	0.624	0.326	1034.46	63.23	0.549
LSTUR + NTD	G-KL	0.282	0.444	0.240	0.612	0.104	0.546	0.828	0.133	0.250	0.754	0.779	0.789	1112.29	30.15	N/A
	PM-2	0.284	0.440	0.244	0.604	0.115	0.546	0.819	0.150	0.250	0.766	0.776	0.789	1112.29	23.86	N/A
	MMR	0.295	0.433	0.235	0.613	0.118	0.575	0.823	0.226	0.270	0.759	0.762	0.807	1112.29	52.36	N/A
	POS	0.365	0.480	0.259	0.667	0.110	0.731	0.844	0.759	0.853	0.699	0.430	0.344	1112.29	212.40	N/A
	ATT	0.371	0.476	0.258	0.667	0.113	0.743	0.840	0.754	0.850	0.712	0.436	0.350	1112.29	212.71	N/A
NPA + NTD	G-KL	0.292	0.469	0.237	0.589	0.108	0.545	0.810	0.133	0.250	0.775	0.779	0.789	1656.96	30.15	N/A
	PM-2	0.302	0.471	0.239	0.576	0.122	0.547	0.806	0.150	0.250	0.783	0.776	0.789	1656.96	23.86	N/A
	MMR	0.314	0.463	0.236	0.598	0.110	0.575	0.808	0.219	0.270	0.772	0.763	0.807	1656.96	52.36	N/A
	POS	0.370	0.487	0.263	0.656	0.108	0.727	0.841	0.755	0.863	0.707	0.431	0.327	1656.96	212.40	N/A
	ATT	0.377	0.484	0.259	0.655	0.111	0.742	0.841	0.742	0.860	0.708	0.448	0.331	1656.96	212.71	N/A
NRMS + NTD	G-KL	0.307	0.482	0.239	0.608	0.091	0.544	0.798	0.133	0.250	0.790	0.779	0.789	1034.46	30.15	N/A
	PM-2	0.316	0.483	0.244	0.598	0.099	0.544	0.796	0.150	0.250	0.794	0.776	0.789	1034.46	23.86	N/A
	MMR	0.315	0.477	0.238	0.616	0.099	0.571	0.798	0.204	0.270	0.790	0.767	0.807	1034.46	52.36	N/A
	POS	0.366	0.480	0.263	0.657	0.110	0.731	0.841	0.765	0.851	0.701	0.419	0.344	1034.46	212.40	N/A
	ATT	0.371	0.476	0.260	0.655	0.112	0.742	0.839	0.747	0.848	0.709	0.444	0.349	1034.46	212.71	N/A
D-RDW	N/A	0.374	0.407	0.229	0.394	0.107	0.556	0.798	0.133	0.250	0.810	0.779	0.789	14.31	2.50	0.554
RP^3_β [15]	N/A	0.222	0.415	0.235	0.582	0.080	0.376	0.840	0.755	0.856	0.743	0.439	0.392	14.31	1.28	0.565
RWE-D [42]	N/A	0.256	0.435	0.222	0.443	0.100	0.372	0.857	0.802	0.842	0.735	0.377	0.433	12.02	0.25	0.554
PLD [25]	N/A	0.152	0.459	0.268	0.418	0.038	0.432	0.801	0.687	0.749	0.782	0.534	0.556	N/A	2.30	N/A
EPD [24]	N/A	0.139	0.443	0.207	0.486	0.081	0.505	0.773	0.611	0.667	0.802	0.577	0.610	N/A	0.14	N/A
Random	N/A	0.180	0.461	0.256	0.705	0.054	0.366	0.756	0.634	0.873	0.842	0.564	0.346	N/A	N/A	0.500

Table 3: Overview for *MIND* of the diversity scores for the top 20 news recommendations and AUC scores for predicting user impressions. The values closest to the perfect score are highlighted in green, second closest in blue, and third closest in red.

Model	Re-ranking	Activ.	Cat. Calib.	Comp. Calib.	Frag.	Alt. Voices	Repr.	Cat. Gini	Sent. Gini	Party Gini	Cat. ILD	Sent. ILD	Party ILD	Train. Cost	Rec. Cost	AUC
Perfect (NTD)	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.133	0.250	1.000	0.779	0.789	N/A	N/A	1.000
LSTUR [3]	N/A	0.266	0.620	0.313	0.503	0.051	0.296	0.855	0.593	0.905	0.618	0.585	0.247	624.72	567.42	0.593
NPA [62]	N/A	0.191	0.574	0.329	0.443	0.076	0.280	0.792	0.614	0.854	0.751	0.587	0.323	1204.69	351.65	0.595
NRMS [63]	N/A	0.211	0.559	0.313	0.712	0.071	0.305	0.756	0.554	0.896	0.764	0.615	0.246	644.78	952.58	0.626
LSTUR + NTD	G-KL	0.250	0.619	0.311	0.558	0.097	0.408	0.786	0.133	0.250	0.769	0.779	0.789	624.72	60.88	N/A
	PM-2	0.227	0.607	0.317	0.574	0.086	0.418	0.809	0.133	0.250	0.734	0.779	0.789	624.72	53.36	N/A
	MMR	0.287	0.604	0.334	0.594	0.102	0.471	0.766	0.000	0.000	0.781	0.789	0.842	624.72	125.83	N/A
	POS	0.299	0.631	0.338	0.693	0.120	0.538	0.790	0.777	0.623	0.693	0.393	0.642	624.72	96.26	N/A
	ATT	0.289	0.633	0.340	0.688	0.134	0.544	0.786	0.746	0.618	0.712	0.438	0.652	624.72	97.53	N/A
NPA + NTD	G-KL	0.226	0.585	0.327	0.458	0.098	0.434	0.829	0.133	0.250	0.691	0.779	0.789	1204.69	60.88	N/A
	PM-2	0.251	0.599	0.331	0.465	0.117	0.438	0.858	0.133	0.250	0.651	0.779	0.789	1204.69	53.36	N/A
	MMR	0.320	0.581	0.329	0.475	0.137	0.517	0.837	0.000	0.000	0.691	0.789	0.842	1204.69	125.83	N/A
	POS	0.329	0.636	0.372	0.549	0.173	0.568	0.900	0.745	0.640	0.436	0.431	0.637	1204.69	96.26	N/A
	ATT	0.331	0.639	0.372	0.556	0.171	0.563	0.896	0.740	0.649	0.449	0.439	0.628	1204.69	97.53	N/A
NRMS + NTD	G-KL	0.232	0.563	0.315	0.692	0.075	0.402	0.713	0.133	0.250	0.812	0.779	0.789	644.78	60.88	N/A
	PM-2	0.238	0.560	0.310	0.706	0.079	0.390	0.729	0.133	0.250	0.801	0.779	0.789	644.78	53.36	N/A
	MMR	0.284	0.574	0.322	0.692	0.088	0.468	0.716	0.000	0.000	0.809	0.789	0.842	644.78	125.83	N/A
	POS	0.346	0.605	0.352	0.713	0.117	0.561	0.766	0.729	0.623	0.740	0.435	0.639	644.78	96.26	N/A
	ATT	0.333	0.603	0.352	0.712	0.119	0.567	0.770	0.713	0.613	0.736	0.461	0.649	644.78	97.53	N/A
D-RDW	N/A	0.281	0.557	0.303	0.515	0.103	0.377	0.722	0.133	0.250	0.822	0.779	0.789	83.41	12.03	0.525
RP^3_β [15]	N/A	0.215	0.543	0.312	0.709	0.074	0.308	0.737	0.540	0.904	0.783	0.622	0.230	18.46	1.47	0.532
RWE-D [42]	N/A	0.225	0.583	0.304	0.398	0.102	0.298	0.724	0.364	0.830	0.805	0.715	0.352	29.22	0.62	0.512
PLD [25]	N/A	0.146	0.580	0.345	0.560	0.059	0.336	0.658	0.484	0.795	0.857	0.665	0.403	N/A	1.22	N/A
EPD [24]	N/A	0.274	0.614	0.318	0.446	0.100	0.399	0.726	0.533	0.850	0.823	0.626	0.377	N/A	0.12	N/A
Random	N/A	0.197	0.618	0.314	0.702	0.057	0.301	0.663	0.503	0.897	0.861	0.649	0.251	N/A	N/A	0.498

Table 4: Overview for *NeMig* of the diversity scores for the top 20 news recommendations and AUC scores for predicting user impressions. The values closest to the perfect score are highlighted in green, second closest in blue, and third closest in red.

Model	Re-ranking	Activ.	Cat. Calib.	Comp. Calib.	Frag.	Alt. Voices	Repr.	Cat. Gini	Sent. Gini	Party Gini	Cat. ILD	Sent. ILD	Party ILD	Train. Cost	Rec. Cost	AUC
Perfect (NTD)	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.250	1.000	0.526	0.789	N/A	N/A	1.000
LSTUR [3]	N/A	0.192	0.637	0.415	0.870	0.044	0.545	0.850	0.951	0.752	0.740	0.048	0.517	149.72	185.12	0.535
NPA [62]	N/A	0.206	0.601	0.420	0.787	0.073	0.568	0.905	0.980	0.725	0.577	0.019	0.556	172.79	107.40	0.491
NRMS [63]	N/A	0.220	0.629	0.428	0.730	0.055	0.556	0.864	0.963	0.808	0.707	0.037	0.468	139.52	152.95	0.552
LSTUR + NTD	G-KL	0.282	0.604	0.385	0.758	0.045	0.565	0.904	0.000	0.250	0.618	0.526	0.789	149.72	16.64	N/A
	PM-2	0.288	0.605	0.385	0.763	0.046	0.565	0.902	0.000	0.250	0.622	0.526	0.789	149.72	8.28	N/A
	MMR	0.192	0.617	0.400	0.826	0.046	0.535	0.876	0.497	0.431	0.675	0.396	0.723	149.72	10.00	N/A
	POS	0.535	0.602	0.371	0.651	0.076	0.614	0.898	0.927	0.667	0.619	0.048	0.552	149.72	29.78	N/A
	ATT	0.543	0.605	0.364	0.576	0.078	0.611	0.896	0.967	0.652	0.632	0.022	0.579	149.72	29.78	N/A
NPA + NTD	G-KL	0.275	0.598	0.385	0.686	0.042	0.571	0.910	0.000	0.250	0.572	0.526	0.789	172.79	16.64	N/A
	PM-2	0.285	0.600	0.386	0.693	0.045	0.571	0.906	0.000	0.250	0.582	0.526	0.789	172.79	8.28	N/A
	MMR	0.234	0.594	0.398	0.736	0.056	0.549	0.913	0.499	0.415	0.548	0.395	0.739	172.79	10.00	N/A
	POS	0.540	0.599	0.373	0.664	0.080	0.615	0.902	0.923	0.707	0.608	0.051	0.510	172.79	29.78	N/A
	ATT	0.538	0.603	0.368	0.588	0.090	0.616	0.899	0.967	0.701	0.628	0.022	0.534	172.79	29.78	N/A
NRMS + NTD	G-KL	0.281	0.624	0.382	0.667	0.040	0.576	0.877	0.000	0.250	0.680	0.526	0.789	139.52	16.64	N/A
	PM-2	0.289	0.622	0.382	0.664	0.037	0.575	0.879	0.000	0.250	0.675	0.526	0.789	139.52	8.28	N/A
	MMR	0.219	0.626	0.411	0.703	0.070	0.552	0.870	0.500	0.452	0.685	0.395	0.705	139.52	10.00	N/A
	POS	0.540	0.611	0.367	0.609	0.086	0.620	0.877	0.935	0.716	0.671	0.043	0.510	139.52	29.78	N/A
	ATT	0.546	0.618	0.363	0.521	0.097	0.625	0.870	0.979	0.739	0.693	0.014	0.500	139.52	29.78	N/A
D-RDW	N/A	0.289	0.598	0.382	0.737	0.042	0.554	0.887	0.000	0.250	0.653	0.526	0.789	95.75	1.60	0.550
RP_{β}^{3} [15]	N/A	0.185	0.612	0.408	0.880	0.049	0.545	0.856	0.921	0.761	0.717	0.076	0.508	94.25	0.14	0.448
RWE-D [42]	N/A	0.186	0.633	0.416	0.877	0.055	0.554	0.850	0.943	0.801	0.762	0.055	0.445	94.28	0.10	0.451
PLD [25]	N/A	0.127	0.625	0.417	0.641	0.027	0.526	0.858	0.926	0.792	0.716	0.074	0.466	N/A	0.22	N/A
EPD [24]	N/A	0.218	0.615	0.433	0.579	0.058	0.587	0.876	0.900	0.733	0.660	0.097	0.558	N/A	0.12	N/A
Random	N/A	0.185	0.628	0.412	0.879	0.047	0.541	0.843	0.929	0.772	0.738	0.068	0.489	N/A	N/A	0.498

Models: The experiment includes LSTUR, NRMS, NPA,[19] RP_{β}^{3}, RWE-D, and the norm-aware D-RDW.[20] Furthermore, we used the filtering algorithms PLD and EPD. A random selection of articles (RND) is used as a baseline for comparison. We calculate the top 20 item recommendations for each user, using reference values for list sizes from the literature [23, 57].

The word embeddings for the neural models are GloVe[21] for MIND as well as fastText for both EB-NeRD (Danish) and NeMig (German).[22] We used off-the-shelf models to calculate article similarity for mapping cold items to users for random walks; the sentence transformers in our workflow include RoBERTa[23] for EB-NeRD, MPNet[24] for MIND, and E5[25] for NeMig.

For D-RDW, NTD covers the dimension of political parties and article sentiment. For all datasets, NTD has buckets for articles that mention 1) the majority parties (30%),2) minority parties (15%), 3) majority *and* minority parties (15%), and 4) no political party (40%). To ensure a broad coverage. the value ranges and percentages for the sentiment distribution are $[-1, -0.5)$ (20%), $[-0.5, 0)$ (30%), $[0, 0.5)$ (30%), $[0.5, 1]$ (20%) for E-NeRD and MIND, and $[-1, 0)$ (50%), $[0, 1]$ (50%) for NeMig.[26]

[19]We did not fine-tune the models, but used the hyperparameters from the official repository: https://github.com/recommenders-team/recommenders

[20]We perform 3 hops, with the exception of D-RDW on NeMig using using 5 hops, as the graph is to sparse to give us 20 items with a lower number of hops.

[21]GloVe word vectors: https://nlp.stanford.edu/projects/glove/

[22]fastText word vectors: https://fasttext.cc/docs/en/crawl-vectors.html

[23]RoBERTa model: https://huggingface.co/FacebookAI/roberta-base

[24]MPNet model: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[25]E5 model: https://huggingface.co/intfloat/multilingual-e5-base

[26]For NeMig, we used only positive and negative sentiments, as the data did not allow for a more detailed assessment of sentiment..

Table 5: Comparison of EB-NeRD, MIND, and NeMig in terms of users, average number or articles in a user history, impressions, articles, and unique article categories.

Dataset	Train Users	Test Users	History (AVG)	Imp. (Total)	Art. (Total)	Cat. (Unique)
MIND	49,823	48,592	21.68	7,336,094	65,058	18
EB-NeRD	15,143	15,339	111.72	3,732,517	11,421	23
NeMig	3,242	3,242	5.78	97,232	4,933	26

Re-rankers: We experiment with the three static re-rankers mentioned in Section 3.3, namely Greedy-KL (G-KL), PM-2, and MMR to calculate the top 20 recommendations for each user. All three static re-rankers use the same diversity dimensions and distributions as those defined by our aforementioned target distribution. More precisely, the diversity dimensions include sentiment and political parties, with equal weights given to both dimensions. We added the DAP dynamic re-ranking strategy that simulated a user with strong position preferences (clicking predominantly on the top-most items, POS) and one reflecting attribute preferences for political parties and sentiment (ATT).

Metrics: We use the following metrics from the literature: 1) five diversity metrics for measuring divergence of activation (Activ.), category calibration (Cat. Calib.), complexity calibration (Comp. Calib.), fragmentation (Frag.), alternative voices (Alt. Voices), and representation (Repr.) adopted from the RADio metrics [57] and 2) two *traditional* diversity metrics, namely the Gini coefficient for article category (Cat. Gini), sentiment (Sent. Gini), and political parties (Party Gini) and with intra-list distance for article category (Cat. ILD), sentiment (Sent. ILD), and political parties (Party ILD). We also report on the upper limit for energy cost (in watt seconds) for training (Train Cost) and for calculating recommendations *per user* (Rec. Cost)[27] and the area under the ROC curve (AUC) for accuracy. Table 2 shows the values for the EB-NeRD dataset, Table 3 for MIND, and Table 4 for NeMig.

5 Results and Discussion

Splitting the RS pipeline into separate stages allows us to look at the impact of each element: the dataset, the recommender models, and the applied re-ranking techniques on target metrics. In outlining these three parts, we focus on the performance comparisons across the RADio metrics (RADio metrics), traditional metrics (Gini and ILD), accuracy (AUC), and computational cost. For each dataset, we compare the families of approaches: 1) traditional neural models (LSTUR, NPA, NRMS), 2) baseline random walk models ($RP_{\beta}^{3}$and RWE-D), and 3) NTD-optimizing algorithms (D-RDW, PLD, and EPD) and re-rankers.

RADio Metrics: RADio metrics assess RS performance on the basis of a divergence between an individual's recommendations and the overall item pool.[28] For **EB-NeRD**, we see that target distribution-optimizing models achieve the top score in all but one category. The performance increase over neural models and re-ranking is especially large for fragmentation and alternative voices. D-RDW creates the least fragmented readership and PLD gives the most exposure to minority positions. A similar picture presents itself with **MIND**. But now we see that NRMS with dynamic position- and topic-aware re-ranking can score higher in terms of activation and representation. The comparatively larger item pool of MIND (it contains more items than users, see Table 1) provides the dynamic approach with sufficient items to diversify the recommendations. This holds true for **NeMig** as well. Dynamic re-ranking for LSTUR, NPA, and NRMS outperforms most other approaches. We also see that the neural models have no top results. With NeMig being more than one order of magnitude smaller than EB-NeRD and MIND in terms of impressions, there does not seem to be enough data available. This also impacts the random walk models, as a low number of impressions means a sparsely connected graph, explaining their poor performance.

Traditional Metrics: The Gini coefficient assesses equality. The smaller the value, the more equal the distribution of a given attribute within the set. ILD measures the average pairwise similarity between items using cosine distance. The smaller the distance, the more similar the items within a list (with respect to a given target dimension).[29] We see that the distribution-optimizing approaches with NTD (G-KL re-ranking and D-RDW in particular) consistently achieves perfect scores for sentiment and party Gini as well as ILD across **all datasets**. The same does not hold true for category Gini/ILD. Part of the reason is that the underlying NTD of D-RDW and the re-rankers do not include the category. On the one hand, this is evidence for the effectiveness of NTD-based approaches. On the other hand, this shows that traditional metrics are but an imperfect solution to capturing diversity.

Accuracy and Computational Cost. We calculate AUC to show that the target distribution-optimizing models and re-rankers not only diversify the recommendation list but also present relevant items similar to the state-of-the-art neural models.[30] For **EB-NeRD**, we see a tie in AUC for neural models and random walks. On **MIND**, neural models substantially outperform the other two families.[31] And the top spots for **NeMig** are shared between NRMS, D-RDW, and LSTUR. While AUC scores around 0.5-0.6 indicated random to poor discriminatory performance, these values reproduce previous findings (cf. [29, 64]), as the news is a particularly difficult domain (containing large shares of cold items and users). In contrast to the AUC performance, we do see normative models and filter algorithms clearly outperforming the re-ranking and neural models in terms of overall energy consumption across all datasets.

6 Limitations and Future Work

Our experiments are limited to offline news recommender systems benchmarking. We chose to present experimental results in the news domain as this allowed us to exemplify the target distribution-optimizing capabilities of our diversity-aware framework extension. Future work could focus on experimenting with different target distributions in other domains. In addition, we need to run online user studies to properly assess the impact of visualization (i.e., article position through varying item placement and accessibility through varying text complexity) on item consumption and engagement of diverse recommendations.

7 Conclusion

We present Informfully Recommenders, a reproducibility framework for recommender algorithms that facilitates diversity-driven offline benchmarking as well as online user experiments. It provides a customizable end-to-end pipeline that allows for seamless algorithm development, benchmarking, and deployment. Targeting user experiments, the pipeline includes text augmentation functionality, lightweight recommendation models, static and dynamic re-rankers, user simulators, normative and traditional diversity metrics, and item visualization. Informfully Recommenders is a modular, diversity-driven extension to the well-established Cornac reproducibility framework

We hope this framework enables researchers to incorporate social norms and values into their algorithms for production-ready recommenders to conducting user studies, as they present the ultimate test for any recommender system [31]. While we illustrated Informfully Recommenders in the news domain, it supports the development, benchmarking, *and* deployment of end-to-end pipelines across different RS domains.

[27] Costs are estimated based on the power consumption of the CPU/GPU and factoring in the overall task runtime [56].

[28] We refer to the original paper [57] for the outline/interpretation of the metrics.

[29] Both Gini and ILD are frequently used as proxies for diversity [34]. However, when looking, e.g., at normative interpretation of exposure diversity (where every political party should get exposure equal to their respective party size), both equality and distance do not adequately represent these targets. We nevertheless report on these values, as they are standard metrics that allow for a comparison with other works.

[30] Please note that PLD and EPD do not make any item classification, therefore AUC is not applicable. Furthermore, the re-rankers do not change the AUC of the underlying model, and we therefore omitted listing the results again.

[31] These models were developed as part for the MIND challenge submission [64]

References

[1] Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1160–1163.

[2] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 336–345.

[3] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 336–345.

[4] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2405–2414.

[5] Andreas Argyriou, Miguel González-Fierro, and Le Zhang. 2020. Microsoft recommenders: Best practices for production-ready recommendation systems. In *Companion Proceedings of the Web Conference 2020*. 50–51.

[6] Christine Bauer, Chandni Bagchi, Olusanmi A Hundogan, and Karin van Es. 2024. Where are the values? a systematic literature review on news recommender systems. *ACM Transactions on Recommender Systems* 2, 3 (2024), 1–40.

[7] Joeran Beel and Haley Dixon. 2021. The 'unreasonable'effectiveness of graphical user interfaces for recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 22–28.

[8] Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Lucien Heitz, Suzanne Tolmeijer, et al. 2021. Diversity in News Recommendation. *Dagstuhl Manifestos* 9, 1 (2021), 43–61.

[9] Keith Bradley and Barry Smyth. 2001. Improving Recommendation Diversity. https://api.semanticscholar.org/CorpusID:11075976

[10] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[11] Pablo Castells, Neil Hurley, and Saul Vargas. 2021. Novelty and diversity in recommender systems. In *Recommender systems handbook*. Springer, 603–646.

[12] Chong Chen, Min Zhang, Yongfeng Zhang, Yiqun Liu, and Shaoping Ma. 2020. Efficient neural matrix factorization without sampling for recommendation. *ACM Transactions on Information Systems (TOIS)* 38, 2 (2020), 1–28.

[13] Xiong-Hui Chen, Bowei He, Yang Yu, Qingyang Li, Zhiwei Qin, Wenjie Shang, Jieping Ye, and Chen Ma. 2023. Sim2rec: A simulator-based decision-making approach to optimize real-world long-term user engagement in sequential recommender systems. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 3389–3402.

[14] Patrick John Chia, Jacopo Tagliabue, Federico Bianchi, Chloe He, and Brian Ko. 2022. Beyond ndcg: behavioral testing of recommender systems with reclist. In *Companion Proceedings of the Web Conference 2022*. 99–104.

[15] Fabian Christoffel, Bibek Paudel, Chris Newell, and Abraham Bernstein. 2015. Blockbusters and wallflowers: Accurate, diverse, and scalable recommendations with random walks. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 163–170.

[16] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 659–666.

[17] Elisia L Cohen. 2002. Online journalism as market-driven journalism. *Journal of broadcasting & Electronic media* 46, 4 (2002), 532–548.

[18] Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 65–74.

[19] Michael D Ekstrand. 2020. Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2999–3006.

[20] Naieme Hazrati and Francesco Ricci. 2022. Simulating users' interactions with recommender systems. In *Adjunct proceedings of the 30th acm conference on user modeling, adaptation and personalization*. 95–98.

[21] Lucien Heitz. 2023. Classification of Normative Recommender Systems. In *Proceedings of the First Workshop on the Normative Design and Evaluation of Recommender Systems*.

[22] Lucien Heitz, Julian A Croci, Madhav Sachdeva, and Abraham Bernstein. 2024. Informfully - Research Platform for Reproducible User Studies. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 660–669.

[23] Lucien Heitz, Oana Inel, and Sanne Vrijenhoek. 2024. Recommendations for the Recommenders: Reflections on Prioritizing Diversity in the RecSys Challenge. In *Proceedings of the Recommender Systems Challenge 2024*. 22–26.

[24] Lucien Heitz, Juliane A Lischka, Rana Abdullah, Laura Laugwitz, Hendrik Meyer, and Abraham Bernstein. 2023. Deliberative Diversity for News Recommendations: Operationalization and Experimental User Study. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 813–819.

[25] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism* 10, 10 (2022), 1710–1730.

[26] Lucien Heitz, Nicolas Mattis, Oana Inel, and Wouter van Atteveldt. 2024. IDEA – Informfully Dataset with Enhanced Attributes. In *Proceedings of the Second Workshop on the Normative Design and Evaluation of Recommender Systems*.

[27] Natali Helberger. 2019. On the democratic role of news recommenders. *Digital Journalism* 7, 8 (2019), 993–1012.

[28] Natali Helberger, Kari Karppinen, and Lucia D'acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.

[29] Andreea Iana, Mehwish Alam, Alexander Grote, Nevena Nikolajevic, Katharina Ludwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023. NeMig-A Bilingual News Collection and Knowledge Graph about Migration. In *Proceedings of the International Workshop on News Recommendation and Analytics 2023*.

[30] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847* (2019).

[31] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara fallacy: Towards more impactful recommender systems research. *Ai Magazine* 41, 4 (2020), 79–95.

[32] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024*. 1–11.

[33] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. RecSys Challenge 2024: Balancing Accuracy and Editorial Values in News Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1195–1199.

[34] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems* 123 (2017), 154–162.

[35] Jiayu Li, Hanyu Li, Zhiyu He, Weizhi Ma, Peijie Sun, Min Zhang, and Shaoping Ma. 2024. ReChorus2. 0: A Modular and Task-Flexible Recommendation Library. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 454–464.

[36] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[37] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The unified framework of media diversity: A systematic literature review. *Digital Journalism* 8, 5 (2020), 605–642.

[38] Pasquale Lops, Marco Polignano, Cataldo Musto, Antonio Silletti, and Giovanni Semeraro. 2023. ClayRS: An end-to-end framework for reproducible knowledge-aware recommender systems. *Information Systems* 119 (2023), 102273.

[39] Hongyu Lu, Min Zhang, Weizhi Ma, Ce Wang, Feng Xia, Yiqun Liu, Leyu Lin, and Shaoping Ma. 2019. Effects of user negative experience in mobile news streaming. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 705–714.

[40] Lien Michiels, Robin Verachtert, and Bart Goethals. 2022. Recpack: An (other) experimentation toolkit for top-n recommendation using implicit feedback data. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 648–651.

[41] Darryl Ong, Quoc-Tuan Truong, and Hady W Lauw. 2024. Cornac-AB: An Open-Source Recommendation Framework with Native A/B Testing Integration. In *Companion Proceedings of the ACM Web Conference 2024*. 1027–1030.

[42] Bibek Paudel and Abraham Bernstein. 2021. Random walks with erasure: Diversifying personalized recommendations on social and information networks. In *Proceedings of the Web Conference 2021*. 2046–2057.

[43] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2016. Updatable, accurate, diverse, and scalable recommendations for interactive applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–34.

[44] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems* 34 (2021), 25944–25955.

[45] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-aware recommender systems. *ACM computing surveys (CSUR)* 51, 4 (2018), 1–36.

[46] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*. 521–530.

[47] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5.

[48] Holli Sargeant, Eliska Pirkova, Matthias C Kettemann, Marlena Wisniak, Martin Scheinin, Emmi Bevensee, Katie Pentney, Lorna Woods, Lucien Heitz, Bojana Kostic, et al. 2022. *Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual.* Organization for Security and Co-operation in Europe.

[49] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems.* 154–162.

[50] Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. 2022. Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2022), 8206–8226.

[51] Panagiotis Symeonidis, Dmitry Chaltsev, Chemseddine Berbague, and Markus Zanker. 2022. Sequence-aware news recommendations by combining intra-with inter-session user information. *Information Retrieval Journal* 25, 4 (2022), 461–480.

[52] Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. 2022. Being Diverse is Not Enough: Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization.* 222–233.

[53] Quoc-Tuan Truong, Aghiles Salah, and Hady Lauw. 2021. Multi-modal recommender systems: Hands-on exploration. In *Fifteenth ACM Conference on Recommender Systems.* 834–837.

[54] Quoc-Tuan Truong, Aghiles Salah, Thanh-Binh Tran, Jingyao Guo, and Hady W Lauw. 2021. Exploring Cross-Modality Utilization in Recommender Systems. *IEEE Internet Computing* (2021).

[55] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems.* 209–216.

[56] Tobias Vente, Lukas Wegmeth, Alan Said, and Joeran Beel. 2024. From clicks to carbon: The environmental toll of recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems.* 580–590.

[57] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio–Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems.* 208–219.

[58] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. 2024. Diversity of what? On the different conceptualizations of diversity in recommender systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.* 573–584.

[59] 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.

[60] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Nava Tintarev, and Jordi Viader Guerrero. 2023. Normalize: The first workshop on normative design and evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems.* 1252–1254.

[61] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (2023), 1–27.

[62] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* 2576–2584.

[63] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP).* 6389–6394.

[64] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 3597–3606.

[65] Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. 2021. Measuring recommender system effects with simulated users. *arXiv preprint arXiv:2101.04526* (2021).

[66] Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. 2023. KuaiSim: A comprehensive simulator for recommender systems. *Advances in Neural Information Processing Systems* 36 (2023), 44880–44897.

[67] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *proceedings of the 30th acm international conference on information & knowledge management.* 4653–4664.

[68] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2912–2923.

[69] Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. 2021. Open benchmarking for click-through rate prediction. In *Proceedings of the 30th ACM international conference on information & knowledge management.* 2759–2769.