EvalRS 2023. Well-Rounded Recommender Systems For Real-World Deployments

Federico Bianchi[1], Patrick John Chia[2], Ciro Greco[3], Claudio Pomo[4], Gabriel Moreira[5], Davide Eynard[6], Fahd Husain[7] and Jacopo Tagliabue[3,8]

[1]*Stanford, Palo Alto, USA*
[2]*Coveo, Canada*
[3]*Bauplan, USA*
[4]*Politecnico di Bari, Bari, Italy*
[5]*NVIDIA, Sao Paulo, Brazil*
[6]*mozilla.ai, United Kingdom*
[7]*mozilla.ai, Canada*
[8]*New York University, USA*

Abstract

EvalRS aims to bring together practitioners from industry and academia to foster a debate on rounded evaluation of recommender systems, with a focus on real-world impact across a multitude of deployment scenarios. Recommender systems are often evaluated only through accuracy metrics, which fall short of fully characterizing their generalization capabilities and miss important aspects, such as fairness, bias, usefulness, informativeness. This workshop builds on the success of last year's workshop at CIKM, but with a broader scope and an interactive format.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

The ubiquity of personalized recommendations in various online platforms, from e-commerce to news to social media, has led to a surge of interest in recommender systems (RS) research. The field has grown to accommodate the new application scenarios, resulting in a plethora of algorithmic approaches to address modelling challenges [1, 2]. However, evaluating RS performance is still difficult, especially considering the increasing complexity of deployments and variety of use cases [3, 4]. In this light, our workshop focuses on multi-dimensional and multi-faceted evaluation techniques: while accuracy metrics are often seen as the proxy for generalization, they miss other important dimensions of real-world systems, such as fairness, informativeness, transparency, and resource constraints [5].

By expanding evaluation beyond traditional accuracy metrics, we aim to better understand the holistic performance of RS across diverse scenarios. The workshop gives participants an in-depth view of multi-dimensional evaluation techniques, allowing them to acquire fundamental skills as well as "live and breath" the problem through a novel format, our *hackathon* (Section 2.1).

2nd Edition of EvalRS: a Rounded Evaluation of Recommender Systems, August 6 - August 10, 2023, Long Beach, CA, USA

fede@stanford.edu (F. Bianchi); ciro.greco@bauplanlabs.com (C. Greco); claudio.pomo@poliba.it (C. Pomo); davide@mozilla.ai (D. Eynard); fahd@mozilla.ai (F. Husain)

CEUR Workshop Proceedings (CEUR-WS.org)

As the enthusiasm about the first edition (EvalRS 2022) showed appetite in the community for re-assessing testing practices, we believe it is time for a new, revised, and improved version of EvalRS. As highlighted even in a recent editorial piece in *Nature Machine Intelligence*, RS "benefits can also give rise to challenging ethical issues" [6]: it is clear that both researchers and industry practitioners want to better understand how to avoid negative societal consequences, such as unfair treatment of users, harmful echo chambers, bias, and increasing levels of polarization and misinformation.

A more nuanced approach to evaluation is essential against this backdrop. With the help of new organizers and the support of mozilla.ai, EvalRS 2023 will help bring these important themes front and center within our community and spark a lively debate on the oldest question of all: how do we know we are doing the right thing?

1. Workshop Description

EvalRS has set out to accomplish a significant goal: to foster closer partnerships between the academic and industrial sectors regarding RS. This is achieved by placing a strong emphasis on comprehensive evaluation techniques that can be effectively applied across diverse domains where RS are utilized. The traditional approach to RS evaluation has been centered on accuracy metrics [7, 8, 9, 10, 11]. However, EvalRS seeks to expand the scope of evaluation techniques beyond just accuracy, to encompass other vital aspects such as fairness, interpretability, and robustness. The ultimate aim of EvalRS is to promote transparency and accountability in the development and deployment of RS by encouraging ethical evaluation metrics that prioritize end-users' interests.

This year's workshop aims to expand on last year's successful event at CIKM. EvalRS2022 was organized as an open source competition (>150 participants in 50 teams) and a popular half-day workshop (>50 attendees). All the materials from the workshop have been released to the public, including accepted papers, models, tests, tutorials, and keynote.[1] The event was a first-of-its-kind initiative and a retrospective has been recently published in Nature Machine Intelligence [12].

Building on the experience and expertise of our diverse group of organizers and PC members, this upcoming edition of our program will be even more comprehensive than before. We are committed to maintaining the open-source spirit that was a hallmark of the first edition while introducing a novel interactive element - a hackathon on evaluation - to make this workshop even more distinctive.

The hackathon will provide participants with a unique opportunity to put their evaluation skills to the test and collaborate with others in real-time. Participants will have the opportunity to work together to solve practical evaluation challenges and gain valuable hands-on experience in the process. The hackathon will serve as an excellent platform for networking and sharing knowledge with other experts in the field.

[1] See https://github.com/RecList/evalRS-CIKM-2022/blob/main/README_CIKM_2022.md for all the links.

1.1. Workshop Theme

EvalRS aims to foster a debate on the rounded evaluation of RS [13, 1], with a focus on real-world impact across a multitude of deployment scenarios. By bringing together experts from industry, academia, and government, EvalRS creates a forum for discussion and collaboration on the latest trends and challenges across a wide range of domains. The main themes from EvalRS 2022 – slice-based metrics, fairness assessment, and the use of representational learning to scale behavioral tests – will be expanded upon in this edition, with particular attention to the social impact of RS [14, 15]. For these reasons, we expect EvalRS 2023 to attract a broad range of practitioners, reflecting the horizontal nature of evaluation challenges: taking last year as an example, participant affiliations ranged from academia (e.g. Politecnico di Milano) to startups (Coveo), from Big Tech (Microsoft) to traditional corporations (Fidelity Investment).

The importance of developing better testing for any deployed system can hardly be overstated, let alone for systems as ubiquitous as RS. We believe that the rounded evaluation of RS is, by nature, a multi-faceted and multi-disciplinary endeavor and that the field as a whole has often been held back by the false dichotomy of *quantitative-and-scalable* vs *qualitative-and-manual* [5]. This workshop can potentially drive innovation and advancement in IR and adjacent fields through our commitment and experience in bridging the gap between industry and academia.

1.2. Call for Papers

We encourage the submission of original contributions along our main topics. Submitted papers will be evaluated (single-blind) according to their originality, technical content, style, clarity, and relevance to the workshop. Papers must be original work and may not be under submission to another venue at the time of review. Accepted papers will appear in the **workshop proceedings**[2].

As a non-exhaustive list, we encourage submissions of long research and position papers (up to 8 pages), short research and position papers (up to 4 pages + refs) and long abstracts (up to 2 pages + refs) on the following topics:

- Online vs offline evaluation - e.g. making offline evaluation more trustworthy and unbiased;
- Tools and frameworks for the evaluation of RS;
- Empirical studies on the evaluation of RS;
- Reports from real-world deployments - failures, successes, and surprises;
- New metrics and methodologies for evaluation, both quantitative and qualitative;
- Multi-dimensional evaluation, combining multiple recommendation quality factors;
- Multi-disciplinary investigation on ethical questions connected to the deployment and use of RS.

Selected papers will be presented with lightning talks during the workshop.

[2]We plan on publishing on CEUR, as we did for EvalRS 2022.

2. Format and Duration

Our workshop will feature a unique interactive event in the form of a hackathon. This will allow participants to collaborate, exchange ideas, and build skills in a supportive and engaging environment. We will provide all the necessary materials and support for the hackathon, and encourage participants to work in teams and submit their projects for consideration.

Aside from the hackathon, we will have a more traditional presentation session where participants can share their research. Additionally, two keynote speakers will share their insights and experiences related to the workshop's topic. We understand that the workshop's value extends beyond the event itself. Therefore, we will release all workshop artifacts, including video recordings and hackathon materials, in an open and accessible format. This will allow participants to revisit and learn from the workshop even after its conclusion, and we hope it will contribute to advancing research in the field.

2.1. Interactive Activity: Hackathon

Our hackathon has been organised as a half-day activity, taking place after the presentations of accepted papers. We will ask participants to come up with a contribution for the rounded evaluation of RS, leveraging an agreed-upon dataset, open-source code, and tools prepared in advance by the organizers. Contribution details will be intentionally left open-ended, as we would like participants to engage different angles of the problem on a shared set of resources. Examples could be operationalizing important notions of robustness, applying and discussing metric definitions from literature, quantifying the trade-off between privacy and accuracy, and so on. The hackathon is a unique opportunity to "live and breathe" the workshop themes, increase chances of multi-disciplinary collaboration, network and discover related work by peers, and contribute valuable materials back to the community.

We plan to release materials in advance of the workshop so that participants can prepare. The event will be open to all the attendees, but we will ask them to register in advance to better prepare the event – for example, we may want to create teams ahead of the event to maximize their diversity.

We plan to award monetary prizes to different contributions in different categories: as the contributions are open-ended, so are the criteria. Generally, we lean towards awarding innovative methodologies and clever ideas, with particular attention to real-world impact: new metrics, interesting model analyses, and thorough qualitative evaluations are all in scope.

Datasets We will be working in a **Two-Sided Digital Aggregator** scenario. We reconciled tracks from the EvalRS2022 dataset with those in the WASABI dataset [16, 17], significantly augmenting the information we had about them. WASABI provides, for a portion of the tracks in the EvalRS dataset (coverage is 48%, for a total of 361218 unique songs), extra features like valence-arousal predictions, emotion and social tags from last.fm, labels determining whether a song is a classic or not, and topic distributions obtained with a LDA topic model (see [18] for an in-depth description of these features). Furthermore, we have used Sentence-Bert [19] with `all-mpnet-base-v2` general-purpose model on WASABI song lyrics to enrich our dataset with song embeddings, calculated both on the full lyrics and on individual verses. The final result is

a dataset with rich metadata and strong baselines for models and tests, ideal to investigate both sides of RS in users and items.

2.2. Sponsorship

2.2.1. Workshop

mozilla.ai will provide sponsorship for the event: the funds will be used to award prizes for important and original work, support students and practitioners from unprivileged backgrounds, and help with the relevant travel expenses (speakers, organizers, etc.).

2.2.2. Social event

Bauplan, Snap, Costanoa Venture will generously provide support for the prize ceremony and the social gathering that will happen after the workshop. For up-to-date details on the logistics, check the workshop website regularly.

3. Tentative Program

3.1. Important Dates

- Paper submissions due: June 16th, 2023
- Paper acceptance notification: June 23th, 2023
- Camera ready deadline: July 6th, 2023
- Workshop day: August 7th, 2023

Participants should refer to the official website (https://reclist.io/kdd2023-cup/) for the latest logistic information on the workshop.

3.2. Keynote Speakers and Program Committee

To reflect a broader view on RS, we are putting together a diverse set of practitioners from **industry** and **academia** as our program committee.

Invited speakers will be chosen from a shortlist of world-renown experts: given the focus on real-world impact, our keynote choice will mostly reflect the importance of ethically relevant deployment scenarios (e.g. misinformation, social networks, biases etc.).

Schedule. Table 1 summarizes our plan for the workshop. Our proposal features a mix of standard event and social gatherings: in particular, our 4 hour workshop during normal KDD hours will be a mix of talks and hackathon (first part). We will then invite participants to stay for a pizza night sponsored by the organizers, and continue the hacking for few more hours: we will then celebrate the winners of the contest at the end of the evening, in a fun and informal environment.

Time	Activity	Details
30 Minutes	**1st Keynote**	First keynote: Luca Belli
30 Minutes	**Papers**	Poster presentations OR lighting talks
2:30 Hours	**Hackathon: part I**	***10 minutes**: materials, rules and the goals.* ***140 minutes**: teamwork*
30 Minutes	**2nd Keynote**	Second keynote: Joey Robinson
Break	**Break**	Break before the pizza night
2 Hours	**After-party, Hackathon: part II**	Pizza night: teamwork, project presentations, prizes

Table 1
Schedule for EvalRS2023

4. Organizing team

Workshop organizers are a mix of academic and industry practitioners, with broad experience in RS and their evaluation. Our team includes the main authors of three popular open-source RS packages for training and evaluation (Merlin[3], Elliot [20], RecList [5]), as well as veterans in the space of Data Challenges and Open Datasets for recommendations (SIGIR 2021 Data Challenge[4], EvalRS 2022[5] [21]).

Federico Bianchi (Stanford) is a postdoctoral researcher at Stanford University. His research, ranging from NLP methods for textual analytics to recommender systems for e-commerce has been published at major NLP and AI conferences (EACL, NAACL, EMNLP, ACL, AAAI, ICLR, RecSys) and journals (Cognitive Science, Applied Intelligence, Semantic Web Journals). Federico co-organized the SIGIR2021 Data Challenge, and the CIKM2022 Data Challenge.
Patrick John Chia (Coveo) is an Applied Scientist at Coveo, focusing on research at the intersection of IR, NLP and eCommerce. He is a co-organizer of multiple data challenges (CIKM 2022, SIGIR 2021) and is a speaker and published author at multiple academic and industry venues (ACL, SIGIR, Berlin Buzzwords). His broad interests lie in better understanding AI methods of today and developing AI with more human-like learning capabilities.
Jacopo Tagliabue (New York University) is co-creator of *RecList*. He is Adj. Professor of MLSys at NYU, speaks regularly at top-tier conferences (including NAACL, WWW, RecSys, SIGIR, KDD), and has served as a committee member for ECNLP, ECONLP, EMNLP, SIRIP, ACL. Jacopo is co-organizer of SIGIR eCom, and was the lead organizer of the SIGIR Data Challenge in 2021, and the CIKM Data Challenge in 2022.
Ciro Greco (Bauplan) holds a Ph.D. in Linguistics and Cognitive Neuroscience at Milano-Bicocca and was a post-doctoral fellow at Ghent University. In 2017, he founded Tooso, which was acquired in 2019 by Coveo. He published extensively in top-tier conferences (including NAACL, ACL, RecSys, SIGIR) and scientific journals (Cognitive Science, Nature Communications). He was also co-organizer of the SIGIR DC 2021 and the CIKM DC in 2022.
Claudio Pomo (Politecnico di Bari) is a research fellow at Politecnico di Bari. His research concerns the issues of responsible AI for personalization, with a particular interest in results'

[3]https://github.com/NVIDIA-Merlin/Merlin
[4]https://sigir-ecom.github.io/ecom2021/data-task.html
[5]https://reclist.io/cikm2022-cup/

reproducibility and multi-objective performance evaluation. Contributions on these topics have been accepted in influential area conferences (SIGIR, RecSys, ECIR, UMAP) and journals (Information Science, IPM). Claudio is among the authors and contributors of Elliot.
Gabriel Moreira (NVIDIA) holds a Ph.D on RS at ITA, Brazil. He is a Sr. Research Scientist and Engineer at NVIDIA Merlin team. He is recognized as a Google Developer Expert (GDE) for Machine Learning since 2019, was a co-organizer of the CIKM Data Challenge 2022, committee member for RecSys and SIGIR conferences and is a distinguished reviewer for ACM TORS journal. He was a member of the teams that won recent RS competitions.
Davide Eynard (mozilla.ai) is a Staff MLE and researcher at mozilla.ai, previously Twitter and Fabula AI. He has been a researcher and lecturer at Università della Svizzera Italiana (USI) and Politecnico di Milano. His research deals with knowledge representation, large-scale multimedia retrieval, computer vision, graph learning, and federated social networks. He has served as a committee member of SIGGRAPH, ECCV, ICCV, CVPR, ICML, NeurIPS, ICLR.
Fahd Husain (mozilla.ai) is the Director of ML at mozilla.ai. He has worked on DARPA research efforts to counter human trafficking, identify narrative propagation on social media, and navigate biomedical graphs for pandemic research. Most recently, he was Principal Investigator for DARPA's World Modelers and Automated Scientific Knowledge Extraction programs. His research is on graph machine learning, weak supervision, and human-in-the-loop paradigms.

Acknowledgments

FB is supported by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence.

References

[1] D. Jannach, P. Pu, F. Ricci, M. Zanker, Recommender systems: Past, present, future, AI Mag. 42 (2021) 3–6.
[2] J. Tagliabue, You do not need a bigger boat: Recommendations at reasonable scale in a (mostly) serverless and open stack, in: RecSys, ACM, 2021, pp. 598–600.
[3] D. Jannach, G. de Souza Pereira Moreira, E. Oldridge, Why are deep learning models not consistently winning recommender systems competitions yet?: A position paper, in: RecSys Challenge, ACM, 2020, pp. 44–49.
[4] K. Higley, E. Oldridge, R. Ak, S. Rabhi, G. de Souza Pereira Moreira, Building and deploying a multi-stage recommender system with merlin, in: RecSys, ACM, 2022, pp. 632–635.
[5] P. J. Chia, J. Tagliabue, F. Bianchi, C. He, B. Ko, Beyond NDCG: behavioral testing of recommender systems with reclist, in: WWW (Companion Volume), ACM, 2022, pp. 99–104.
[6] Editors, Algorithmic recommendations, anyone?, Nature Machine Intelligence (2023). doi:10.1038/s42256-023-00631-7.
[7] F. Bianchi, J. Tagliabue, B. Yu, Query2prod2vec: Grounded word embeddings for ecommerce, in: NAACL-HLT (Industry Papers), Association for Computational Linguistics, 2021, pp. 154–162.

[8] P. Kouki, I. Fountalis, N. Vasiloglou, X. Cui, E. Liberty, K. A. Jadda, From the lab to production: A case study of session-based recommendations in the home-improvement domain, in: RecSys, ACM, 2020, pp. 140–149.
[9] A. Rashed, S. Jawed, L. Schmidt-Thieme, A. Hintsches, Multirec: A multi-relational approach for unique item recommendation in auction systems, in: RecSys, ACM, 2020, pp. 230–239.
[10] 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: RecSys, ACM, 2020, pp. 23–32.
[11] V. W. Anelli, A. Bellogín, T. D. Noia, D. Jannach, C. Pomo, Top-n recommendation algorithms: A quest for the state-of-the-art, in: UMAP, ACM, 2022, pp. 121–131.
[12] J. Tagliabue, F. Bianchi, T. Schnabel, G. Attanasio, C. Greco, G. de Souza Moreira, P. J. Chia, A challenge for rounded evaluation of recommender systems, Nature Machine Intelligence (2023) 1–2.
[13] P. Cremonesi, D. Jannach, Progress in recommender systems research: Crisis? what crisis?, AI Mag. 42 (2021) 43–54.
[14] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, AI Soc. 35 (2020) 957–967.
[15] M. Elahi, D. K. Kholgh, M. S. Kiarostami, S. Saghari, S. P. Rad, M. Tkalcic, Investigating the impact of recommender systems on user-based and item-based popularity bias, Inf. Process. Manag. 58 (2021) 102655.
[16] M. Fell, E. Cabrio, M. Tikat, F. Michel, M. Buffa, F. Gandon, The WASABI Song Corpus and Knowledge Graph for Music Lyrics Analysis, Language Resources and Evaluation (2022). URL: https://hal.science/hal-03812106.
[17] M. Buffa, E. Cabrio, M. Fell, F. Gandon, A. Giboin, R. Hennequin, F. Michel, J. Pauwels, G. Pellerin, M. Tikat, M. Winckler, The WASABI Dataset: Cultural, Lyrics and Audio Analysis Metadata About 2 Million Popular Commercially Released Songs, in: The Semantic Web. ESWC 2021. Lecture Notes in Computer Science, vol 12731., 2021, pp. 515–531. URL: https://hal.science/hal-03282619. doi:10.1007/978-3-030-77385-4_31.
[18] M. Fell, E. Cabrio, E. Korfed, M. Buffa, F. Gandon, Love me, love me, say (and write!) that you love me: Enriching the wasabi song corpus with lyrics annotations, arXiv abs/1912.02477 (2019).
[19] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.
[20] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: SIGIR, ACM, 2021, pp. 2405–2414.
[21] J. Tagliabue, F. Bianchi, T. Schnabel, G. Attanasio, C. Greco, G. de Souza P. Moreira, P. J. Chia, Evalrs: a rounded evaluation of recommender systems, in: CIKM Workshops, volume 3318 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.