Unveiling the Potential of Recommender Systems through Multi-Objective Metrics*

Discussion Paper

Vincenzo Paparella[1,*], Dario Di Palma[1,*], Vito Walter Anelli[1], Alessandro De Bellis[1] and Tommaso Di Noia[1]

[1]*Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy*

Abstract

Current recommender systems (RSs) prioritize accuracy, often neglecting aspects like diversity and fairness. This single-metric approach overlooks valuable trade-offs between different qualities. We propose a multi-objective evaluation using Pareto optimality and ***Quality Indicators*** **(QI)** of Pareto frontiers to consider all model configurations simultaneously across multiple perspectives. This approach reveals a more comprehensive picture of RS performance, potentially leading to a reevaluation of existing methods. Code and data are available at https://github.com/sisinflab/RecMOE.

Keywords

Recommender System, Multi-Objective Evaluation, Pareto optimality

1. Introduction

The success of Recommender Systems (RSs) is often measured by their ability to accurately predict a user's preferences and suggest relevant items. However, beyond-accuracy metrics like diversity [2], novelty [3, 4], and fairness [5, 6] have been proposed. While beyond-accuracy metrics have gained momentum, accuracy is still prioritized [7, 8, 9]. Figure 1 shows the normalized performance of baselines on the Goodreads dataset, selecting the best hyper-parameters for each metric. Selecting the best model solely based on accuracy limits consideration of beyond-accuracy performance. A Pareto-optimal configuration improves at least one objective without hurting others, forming the Pareto frontier [10, 11]. We propose introducing ***Quality Indicators*** (QIs) [12] to RSs, providing a quantitative evaluation of Pareto frontiers from different perspectives [13]. Our contributions are (i) Showing the negative impact of prioritizing accuracy and motivating multi-objective evaluation; (ii) Computing Pareto frontiers for hyper-parameter settings of models on public datasets in multi-objective scenarios. (iii) Enhancing multi-objective evaluation by utilizing QIs to comprehensively analyze recommendation models.

2. Quality Indicators

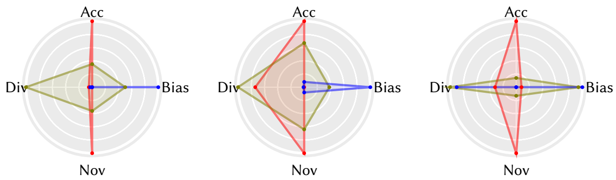
In this Section, we present the Quality Indicators (QIs) to assess the Pareto frontiers corresponding to an RS model. They can be classified according to the quality they assess.

IIR2024: 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy

*Extended version [1] published at the 17th ACM Conference on Recommender Systems (RecSys 2023).

*Corresponding authors.

vincenzo.paparella@poliba.it (V. Paparella); d.dipalma2@phd.poliba.it (D. D. Palma)



(a) UserKNN (b) $RP^3\beta$ (c) $EASE^R$

Models chosen for the best values of — Accuracy/Novelty — Diversity — Bias

Figure 1: Kiviat diagrams indicating the performance of the models on the Goodreads dataset. The models are selected according to different metrics for each objective. Higher means better.

Spread QI. The QIs for Spread indicate the range of the Pareto-optimal solutions on the Pareto frontier. For our study, we use the Maximum Spread (MS) [14]. Specifically, this spread indicator measures the range of a Pareto frontier by considering the maximum extent of each objective. The higher the value, the better the extensiveness of the curve.
Uniformity QI. The uniformity of a Pareto frontier provides information about the distribution of the solutions. A higher uniformity of the curve denotes that the solutions are less dispersed, while a low uniformity indicates more diversity within the set. Specifically, we employ the Spacing metric (SP) [15] that measures the variation in the Manhattan distances between the Pareto-optimal solutions. The lower the value, the more concentrated the solutions are on the Pareto frontier. However, an $SP = 0$ indicates that all the solutions could be equidistant.
Cardinality QI. Given K generic solutions belonging to the set B, the QIs for cardinality determine the proportion of Pareto-optimal solutions in this set. Specifically, the Error Ratio (ER) [16] is defined as $ER(B) = \frac{\sum_{b \in B} e(b)}{K}$ with $e(b) = 1$ if b is a Pareto-optimal solution, 0 otherwise. A higher ER value indicates greater Pareto-optimal solutions in the set B.
All quality aspects QI. The QIs included in this category provide insights into the spread, uniformity, and cardinality of the Pareto frontiers simultaneously. Among them, the Hypervolume (HV) [17] is a volume-based QI that measures the volume of the objective function space dominated by the Pareto frontier. The larger the hypervolume, the better the solution set is.

3. Experiments

We aim to answer two research questions: **RQ1**: *To what extent can the models provide Pareto-optimal configurations? Are these configurations uniformly distributed, or are they dispersed enhancing diverse solutions to the trade-off?* **RQ2**: *Which model has the Pareto frontier that simultaneously offers better solutions on multiple metrics?*

Datasets. We select three different datasets to cover several domains. Specifically, we use *Amazon Music* (music), *Goodreads* [18] (books), and *Movielens1M* [19] (movies).

Baselines and Hyper-parameters Settings Exploration. We train five recommendation algorithms, i.e., $EASE^R$ [20], MultiVAE [21], LightGCN [22], $RP^3\beta$ [23], and UserKNN [24]. We train 32 hyper-parameter values combinations of each model by using Elliot [25].

Metrics. We assess the baselines' performance under several perspectives. We compute nDCG, Precision, and Recall for the accuracy of recommendations. From the final user point of view, we evaluate the diversity (with Gini index [26] and Item Coverage) and novelty (with EPC and

Table 1
QIs of the Pareto frontiers results for the identified scenarios. The arrow indicates the descending or ascending order for the best solution. SP has no specific order of solutions, since its interpretation is strictly connected with the MS indicator. C counts how many solutions lay on the Pareto frontier.

	Model	Objectives									
		Accuracy / Novelty / Diversity					Accuracy / Bias				
		HV↑	ER↑	MS↑	SP	C↑	HV↑	ER↑	MS↑	SP	C↑
Amazon Music	EASE^R	**0.00095**	**0.46875**	0.24986	0.01476	**15**	0.01355	**0.43750**	0.11886	0.00669	**14**
	UserKNN	0.00082	0.34375	**0.29452**	0.00496	11	0.01448	0.34375	**0.17871**	0.00980	11
	LightGCN	0.00051	0.06250	0.01335	0.00000	2	0.00835	0.03125	0.00000	0.00000	1
	MultiVAE	0.00022	0.12500	0.09656	0.01738	4	0.00468	0.15625	0.05629	0.00351	5
	RP$^3\beta$	0.00039	0.18750	0.20753	0.05888	6	**0.03489**	0.21875	0.11336	0.01173	7
Goodreads	EASE^R	0.00074	**0.59375**	0.09910	0.00227	**19**	0.00439	0.65625	0.09433	0.00214	21
	UserKNN	**0.00110**	0.31250	**0.19889**	0.01287	10	0.02267	**0.71875**	**0.48042**	0.01471	**23**
	LightGCN	0.00051	0.18750	0.06743	0.00783	6	0.00696	0.18750	0.09180	0.01536	6
	MultiVAE	0.00043	0.06250	0.05022	0.00000	2	0.00521	0.06250	0.01827	0.00000	2
	RP$^3\beta$	0.00083	0.12500	0.05584	0.01213	4	**0.05544**	0.28125	0.29529	0.02657	9
Movielens1M	EASE^R	0.00865	**0.68750**	0.09833	0.00446	**22**	0.00281	**0.65625**	0.06001	0.00196	**21**
	UserKNN	**0.01296**	0.28125	**0.30929**	0.03641	9	0.08191	0.50000	0.52723	0.01810	16
	LightGCN	0.00807	0.18750	0.01012	0.00287	6	0.00974	0.15625	0.00617	0.00181	5
	MultiVAE	0.01216	0.21875	0.03419	0.00427	7	0.01639	0.18750	0.02528	0.00293	6
	RP$^3\beta$	0.00839	0.06250	0.03796	0.00000	2	**0.14014**	0.46875	**0.86913**	0.03228	15

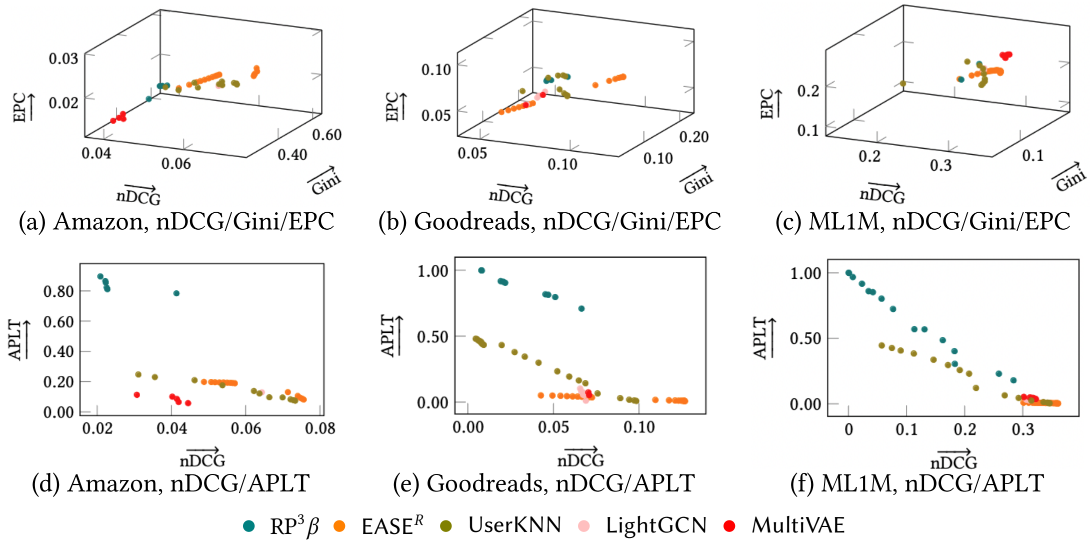
EFD [3]). Finally, we measure the popularity bias of the recommendations with APLT [27] – the greater, the better – and ARP [26] – the less, the better. All these metrics refer to cutoff 10.

Multi-Objective Evaluation Methodology. We obtain Pareto frontiers for each recommender system (RS) baseline using the metrics described in Section 2. Each hyper-parameter setting represents a solution in the objective space. We identify the Pareto-optimal configurations for each baseline, forming their respective Pareto frontiers. We evaluate these frontiers using QIs under two scenarios: 1) user-centered (accuracy, diversity, novelty) and 2) accuracy vs. algorithmic bias. Figure 2 shows the resulting Pareto frontiers.

3.1. Results and Discussion

While EASE^R and UserKNN provide the most accurate recommendations, beyond-accuracy metrics paint a different picture. By observing Figure 2, UserKNN exhibits better diversity than EASE^R. Finally, RP$^3\beta$ consistently outperforms its competitors in addressing the popularity bias. We delve into a multi-objective evaluation using QIs on Pareto frontiers. Here, we examine the distribution of Pareto-optimal configurations and performance on all quality metrics.

Distribution of Pareto-optimal configurations. The Error Ratio (ER), Maximum Spread (MS), and Spacing metric (SP) values in Table 1 unveil interesting insights into the distribution of Pareto-optimal configurations for each model. In the nDCG/APLT scenario for the Movielens1M dataset, for instance: 1) UserKNN exhibits a wide range of solutions with good dispersion across the Pareto frontier, indicating its ability to offer various well-balanced trade-offs between accuracy and algorithmic bias; 2) EASE^R, while offering a high number of solutions on the frontier, they tend to be concentrated in a limited area, suggesting a lack of diversity in the achievable trade-offs; 3) RP$^3\beta$ strikes a good balance between the number of solutions, their dispersion, and the ability to provide various trade-offs between accuracy and bias. This is reflected in its high ER, MS, and SP values. Similar trends are observed for the other datasets



(a) Amazon, nDCG/Gini/EPC (b) Goodreads, nDCG/Gini/EPC (c) ML1M, nDCG/Gini/EPC

(d) Amazon, nDCG/APLT (e) Goodreads, nDCG/APLT (f) ML1M, nDCG/APLT

Figure 2: Pareto optimal solutions plots for Amazon Music, Goodreads, and MovieLens1M. The first row refers to the nDCG/Gini/EPC scenario, and the second row refers to the nDCG/APLT scenario. The arrows indicate the optimal directions.

(see Figures 2f - 2e). When examining the user-centric scenario (nDCG/Gini/EPC), UserKNN again excels, offering well-diversified solutions across all datasets (see Figures 2a - 2c).

Performance on all quality metrics. In response to RQ2, we can utilize the Hypervolume (HV) measure. HV evaluates the performance of models from multiple objectives simultaneously, as shown in Table 1. By considering the cardinality and dispersion of the Pareto-optimal solutions and the dominance among the Pareto frontiers, HV provides us with valuable insights. The higher the volume or area under the frontier, the greater the HV. The results show that UserKNN outperforms the other models by achieving the best or second-best values of HV for all datasets and scenarios. This result indicates that UserKNN generates an extensive and diversified Pareto frontier while performing well across all metrics. While EASE^R has the highest value of HV for the Amazon Music dataset in the user-centred scenario, it does not dominate or get dominated in the remaining cases. This result highlights the model's limited reliance on accounting for multiple metrics. LightGCN shows no distinctive trends, while MultiVAE's HV decreases when dealing with sparser datasets. RP$^3\beta$ confirms its capability in managing the nDCG/APLT trade-off by achieving the highest values of HV and visual dominance of its Pareto frontiers against the others in Figures 2d, 2e, and 2f.

4. Conclusion and Future Work

Our multi-objective evaluation with Quality Indicators reveals new insights into recommender systems (RSs). While EASE^R exhibits high accuracy, UserKNN emerges as a strong contender offering diverse solutions across multiple objectives. Additionally, RP$^3\beta$ proved to be highly effective in the accuracy/algorithmic bias scenario.

Acknowledgements. The authors acknowledge partial support of the following projects: OVS: Fashion Retail Reloaded, Lutech Digitale 4.0, Secure Safe Apulia, Patti Territoriali WP1, BIO-D, and MOST - Centro Nazionale per la Mobilità Sostenibile. We also gratefully acknowledge the CINECA award under the ISCRA initiative, for the availability of HPC resources and support.

References

[1] V. Paparella, D. Di Palma, V. W. Anelli, T. D. Noia, Broadening the scope: Evaluating the potential of recommender systems beyond prioritizing accuracy, in: J. Zhang, L. Chen, S. Berkovsky, M. Zhang, T. D. Noia, J. Basilico, L. Pizzato, Y. Song (Eds.), Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023, ACM, 2023, pp. 1139–1145. URL: https://doi.org/10.1145/3604915.3610649. doi:10.1145/3604915.3610649.

[2] V. Paparella, V. W. Anelli, L. Boratto, T. D. Noia, Reproducibility of multi-objective reinforcement learning recommendation: Interplay between effectiveness and beyond-accuracy perspectives, in: J. Zhang, L. Chen, S. Berkovsky, M. Zhang, T. D. Noia, J. Basilico, L. Pizzato, Y. Song (Eds.), Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023, ACM, 2023, pp. 467–478. URL: https://doi.org/10.1145/3604915.3609493. doi:10.1145/3604915.3609493.

[3] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: B. Mobasher, R. D. Burke, D. Jannach, G. Adomavicius (Eds.), Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011, ACM, 2011, pp. 109–116. URL: https://dl.acm.org/citation.cfm?id=2043955.

[4] D. Di Palma, Retrieval-augmented recommender system: Enhancing recommender systems with large language models, in: RecSys, ACM, 2023, pp. 1369–1373.

[5] L. Boratto, G. Fenu, M. Marras, Interplay between upsampling and regularization for provider fairness in recommender systems, User Model. User Adapt. Interact. 31 (2021) 421–455. URL: https://doi.org/10.1007/s11257-021-09294-8. doi:10.1007/s11257-021-09294-8.

[6] D. Di Palma, V. W. Anelli, D. Malitesta, V. Paparella, C. Pomo, Y. Deldjoo, T. D. Noia, Examining fairness in graph-based collaborative filtering: A consumer and producer perspective, in: IIR, volume 3448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 79–84.

[7] V. W. Anelli, T. D. Noia, E. D. Sciascio, C. Pomo, A. Ragone, On the discriminative power of hyper-parameters in cross-validation and how to choose them, in: T. Bogers, A. Said, P. Brusilovsky, D. Tikk (Eds.), Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019, ACM, 2019, pp. 447–451. URL: https://doi.org/10.1145/3298689.3347010. doi:10.1145/3298689.3347010.

[8] V. W. Anelli, A. Bellogín, T. D. Noia, C. Pomo, Reenvisioning the comparison between neural collaborative filtering and matrix factorization, in: H. J. C. Pampín, M. A. Larson, M. C. Willemsen, J. A. Konstan, J. J. McAuley, J. Garcia-Gathright, B. Huurnink, E. Oldridge (Eds.), RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021, ACM, 2021, pp. 521–529. URL: https://doi.org/10.1145/3460231.3475944. doi:10.1145/3460231.3475944.

[9] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. D. Noia, E. D. Sciascio, Evaluating chatgpt as a recommender system: A rigorous approach, CoRR abs/2309.03613 (2023).

[10] R. Marler, J. Arora, Survey of multi-objective optimization methods for engineering, Structural and Multidisciplinary Optimization 26 (2004) 369–395. doi:10.1007/

s00158-003-0368-6.
[11] V. Paparella, V. W. Anelli, F. M. Nardini, R. Perego, T. D. Noia, Post-hoc selection of pareto-optimal solutions in search and recommendation, in: I. Frommholz, F. Hopfgartner, M. Lee, M. Oakes, M. Lalmas, M. Zhang, R. L. T. Santos (Eds.), Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023, ACM, 2023, pp. 2013–2023. URL: https://doi.org/10.1145/3583780.3615010. doi:10.1145/3583780.3615010.
[12] M. Li, X. Yao, Quality evaluation of solution sets in multiobjective optimisation: A survey, ACM Computing Surveys (CSUR) 52 (2019) 1–38.
[13] V. Paparella, Pursuing optimal trade-off solutions in multi-objective recommender systems, in: J. Golbeck, F. M. Harper, V. Murdock, M. D. Ekstrand, B. Shapira, J. Basilico, K. T. Lundgaard, E. Oldridge (Eds.), RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022, ACM, 2022, pp. 727–729. URL: https://doi.org/10.1145/3523227.3547425. doi:10.1145/3523227.3547425.
[14] E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: Empirical results, Evolutionary computation 8 (2000) 173–195.
[15] J. R. Schott, Fault tolerant design using single and multicriteria genetic algorithm optimization., Technical Report, Air force inst of tech Wright-Patterson afb OH, 1995.
[16] D. A. Van Veldhuizen, Multiobjective evolutionary algorithms: classifications, analyses, and new innovations, Air Force Institute of Technology, 1999.
[17] E. Zitzler, L. Thiele, Multiobjective optimization using evolutionary algorithms—a comparative case study, in: Parallel Problem Solving from Nature—PPSN V: 5th International Conference Amsterdam, The Netherlands September 27–30, 1998 Proceedings 5, Springer, 1998, pp. 292–301.
[18] M. Wan, R. Misra, N. Nakashole, J. J. McAuley, Fine-grained spoiler detection from large-scale review corpora, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 2605–2610. URL: https://doi.org/10.18653/v1/p19-1248. doi:10.18653/v1/p19-1248.
[19] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 19:1–19:19. URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.
[20] H. Steck, Embarrassingly shallow autoencoders for sparse data, in: L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 3251–3257. URL: https://doi.org/10.1145/3308558.3313710. doi:10.1145/3308558.3313710.
[21] D. Liang, R. G. Krishnan, M. D. Hoffman, T. Jebara, Variational autoencoders for collaborative filtering, in: P. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, ACM, 2018, pp. 689–698. URL: https://doi.org/10.1145/3178876.3186150. doi:10.1145/3178876.3186150.
[22] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering

graph convolution network for recommendation, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 639–648. URL: https://doi.org/10.1145/3397271.3401063. doi:10.1145/3397271.3401063.

[23] B. Paudel, F. Christoffel, C. Newell, A. Bernstein, Updatable, accurate, diverse, and scalable recommendations for interactive applications, ACM Trans. Interact. Intell. Syst. 7 (2017) 1:1–1:34. URL: https://doi.org/10.1145/2955101. doi:10.1145/2955101.

[24] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: An open architecture for collaborative filtering of netnews, in: J. B. Smith, F. D. Smith, T. W. Malone (Eds.), CSCW '94, Proceedings of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC, USA, October 22-26, 1994, ACM, 1994, pp. 175–186. URL: https://doi.org/10.1145/192844.192905. doi:10.1145/192844.192905.

[25] 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: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2405–2414. URL: https://doi.org/10.1145/3404835.3463245. doi:10.1145/3404835.3463245.

[26] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, User Model. User Adapt. Interact. 25 (2015) 427–491. URL: https://doi.org/10.1007/s11257-015-9165-3. doi:10.1007/s11257-015-9165-3.

[27] H. Abdollahpouri, R. Burke, B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking, in: R. Barták, K. W. Brawner (Eds.), Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019, AAAI Press, 2019, pp. 413–418. URL: https://aaai.org/ocs/index.php/FLAIRS/FLAIRS19/paper/view/18199.