Annotation Practices in Societally Impactful Machine Learning Applications
What are Popular Recommender Systems Models Actually Trained On?

Sav, Andra Georgiana; Demetriou, Andrew M.; Liem, Cynthia C.S.

Publication date
2023

Document Version
Final published version

Published in
CEUR Workshop Proceedings

Annotation Practices in Societally Impactful Machine Learning Applications: What are Popular Recommender Systems Models Actually Trained On?

Andra-Georgiana Sav[1], Andrew M. Demetriou[1] and Cynthia C. S. Liem[1]

[1]*Delft University of Technology, The Netherlands*

Abstract

Machine Learning (ML) models influence all aspects of our lives. They also commonly are integrated in recommender systems, which facilitate users' decision-making processes in various scenarios, such as e-commerce, social media, news and online learning. Training performed on large volumes of data is what ultimately drives such systems to provide meaningful recommendations. However, a lack of standardized practices has been observed when it comes to data collection and annotation methods for ML datasets. This research paper systematically identifies and synthesizes the state of standardization with regard to data collection and annotation reporting in the recommender systems domain, through a systematic literature view into the 100 most-cited recommender systems papers from the most impactful venues within the Computing and Information Technology field. Multiple facets of the employed techniques are touched upon, such as reported human annotations and annotator diversity, label quality, and the public availability of training datasets. Recurrent use of just a few benchmark datasets, poor documentation practices, and reproducibility issues in experiments are some of the most striking findings uncovered by this study. We discuss the necessity of transitioning from pure reliance on algorithmic performance metrics to prioritizing data quality and fit. Finally, concerns are raised when it comes to biases and socio-psychological factors inherent in the datasets, and further exploration of embedding these early in the design of ML models is suggested.

Keywords

machine learning, recommender systems, data collection, annotation practices, societal impact

1. Introduction

Automated systems are fueled by data—yet when it comes to annotating data for Machine Learning (ML) models, a lack of standardized practices, processes or training of practitioners will affect the reliability of the produced output. As mentioned by [1], most of the current ML research focuses on optimizing accuracy metrics to evaluate the correctness of outputs, rather than establishing qualitative data collection and standardized annotation methods. Recent work [2] further highlights the current challenges when it comes to data annotation practices, pointing out how "practitioners described nuanced understandings of annotator diversity, but rarely designed dataset production to account for diversity in the annotation process".

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2023), September 19th, 2023, co-located with the 17th ACM Conference on Recommender Systems, Singapore, Singapore.

andra-sav@outlook.com (A. Sav); a.m.demetriou@tudelft.nl (A. M. Demetriou); c.c.s.liem@tudelft.nl (C. C. S. Liem)

In order to get more structured and systematic insight into the degree—or, lack—of data annotation practices in societally impactful ML application domains, a topic proposal was submitted to the 3rd year undergraduate Research Project in the Computer Science and Engineering curriculum at Delft University of Technology. Here, students were invited to choose an impactful applied machine learning domain or publication venue of their choice, and conduct a systematic literature review in which annotation practices were documented for the top cited publications for their chosen domain or venue. In this article, we will report on the outcomes of such a review for the domain of recommender systems.

Recommender systems have emerged as powerful instruments in nowadays' society, with use cases spanning across industries (e.g., media, banking, telecom, retail). As [3] notes, Netflix's system design revolves around the idea that "everything is a recommendation". On the same note, major providers such as Google, Amazon, and LinkedIn make use of profiling mechanisms to build and expand their businesses. In their literature review on recommender systems and the ethical challenges they pose, [4] distinguishes six areas of concern, mapping each of them to a possible solution. Within these proposals, one can note the need for introducing factual explanations, as well as increasing the transparency of user categorization to minimize the concerns regarding opacity and lack of user autonomy and personal identity. Here, the interests of account providers and system owners interests (e.g., increase sales of specific products, news propaganda) might not be aligned with the users' original intents of managing information overload, also emphasizing the need for multi-stakeholder perspectives [5].

Often, users are unaware of how recommender systems actually work. If so, they might be misled to believe that the recommendations meaningfully reflect their own interests, while selective exposure to certain categories may steer their future choices towards those, and reshape their personal preferences without them noticing. Therefore, it is important to have accountability on the manner in which recommendations are provided. With much of the technical recommender mechanisms including ML components, this also calls for more thorough understanding of data collection and annotation practices, as these will fundamentally impact any consequent system components.

The central question to this paper is *What are popular recommender systems models actually trained on?*. This will be done by systematically capturing the extent to which the most-cited recommender system papers present in impactful venues within the Computing and Information Technology field have reported explainable data collection and annotation practices, for the purpose of adopting a transparent, fair, and user-centered approach early in the design of the recommendation system. To ensure the scope of the search is clearly defined, the study methodology is further outlined in Section 2. Review results are summarized in Section 3, followed by a Discussion in Section 4. Findings and key implications are briefly summarized in Section 5. Finally, to emphasize our researcher accountability, we reflect on responsible research considerations in Appendix A, and acknowledge individual author contributions in Appendix B.

2. Methodology

The current research paper is based on a systematic review method, as this provides a clear, structured framework "to collect, identify, and critically analyze the available research studies

(e.g., articles, conference proceedings, dissertations) through a systematic procedure" [6]. This method has been initially used to gather relevant information sources, after which the paper progresses to explore and analyze some of the datasets employed by the reviewed papers.

In the upcoming subsections, the methods used to collect data will be explained in accordance to the PRISMA guidelines [7], which is an evidence-based framework commonly adopted when reporting systematic reviews.

2.1. Information sources

All papers reviewed were sampled from the ACM Digital Library, as it is one of the most comprehensive databases in the domain of Computing and Information Technology [8].

2.2. Search strategy

In terms of the search criteria used, only English papers published at most 5 years ago were considered, as to capture the practices in state-of-the-art systems. Moreover, the filtering has been done considering papers having "recommender system(s)"/"recommendation system(s)" in the title, and terms such as "supervised machine learning" or "supervised technique(s)" in the full text. The selection of these specific criteria allows the assessment of current practices in recommender systems that are possibly built with supervised learning (but not limited to it as the only technique). The search strings were run on May 8, 2023. From the resulting papers, the top-100 most cited papers are more thoroughly reviewed. This is based on the need to narrow down the research to a feasible scope[1], while it also is a good indicator of the current practices within the papers that create the most impact within this field. The resulting set of papers encompasses [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 60, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107].

2.3. Data collection process

The data collection has entirely been done by the first author of this paper, who firstly examined the abstract of the paper to check its relevance. Then, the full text was scanned to check for mentions of data collection and annotation practices. More specifically, inspired by the procedure adopted by [1], for each paper the following sub-questions were answered:

1. Was the work an original task?
2. Did the work use human annotations as labels for the training data?
3. Were original human annotations (i.e., annotations collected by themselves) or external human annotations used (i.e., annotations from an existing dataset)?
4. Who were the annotators? (i.e., what population were they drawn from?)
5. Was the number of annotators specified?
6. Was the number of annotators estimated beforehand?

[1]The course assignment ran over a period of 10 weeks.

Table 1
Topic diversity of reviewed papers.

Topic	Mentioned by
Deep-learning based Recommender Systems	14
Graph-based Recommender Systems	11
Conversational Recommender Systems	10
Attacks to Recommender Systems	4
Knowledge-Distillation in Recommender Systems	4

7. Were there formal instructions for the annotators?
8. Was there a required training for the annotators?
9. Was there any pre-screening done for the annotators on the crowdwork platforms?
10. Did multiple annotators label the same item?
11. Was there any reported inter-annotator agreement?
12. Was there any metric specifying the label quality?
13. Was a link to the dataset provided?

The unavailability of data is also taken into consideration, as the ultimate goal is to establish to which extent the authors explicitly mention the data collection or annotation practices. Collection of data is done systematically. Results of a paper analysis are immediately noted down in the results table[2]. In case an answer to a sub-question is uncertain, this is explicitly noted as well, to mitigate any possible error of judgment. Where provided, links to the datasets used and their referencing paper are stored for further analysis. Additionally, papers excluded from the review process are also stored, and a short explanation for doing so is given[3]. A double-check of each entry is performed before proceeding.

2.4. Data overview

This section intends to provide more in-depth insights into the collected data, as to guide an accurate interpretation of the results discussed later. Hence, the reviewed literature has been further categorized into several aspects. With regard to **publication year**, as depicted in Figure 1, more than half of the literature under review was published in 2020 or 2021, with less than 10% being published in 2022. Furthermore, as for **topic diversity**, we identified several recurrent paper topic categories, as presented in Table 1. It is important to note that the enumerated categories are not exhaustive; instead, they serve as an overview of recurring themes identified during the review process. As specific datasets could exhibit greater suitability for particular scenarios, it may prove useful to bear these categories in mind when assessing the distribution of datasets used.

[2]Full results are published online at https://airtable.com/shrP0DCwzaMVdJRsA.
[3]This information is accessible in a separate "Excluded Papers" table at https://airtable.com/shrbq6E0DxSo82rCo.

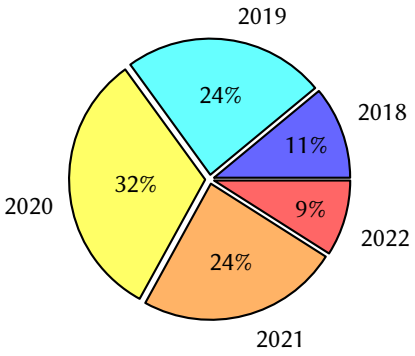

Figure 1: Reviewed papers distribution based on publication year.

3. Findings

This section summarizes the main findings with regard to the established research sub-questions previously mentioned in Section 2.3. We report results on the originality of tasks encountered in the papers, use of human annotations (either internal or external), details regarding the annotators and annotation process, as well as label quality and links to datasets.

Task originality. Given the complexity and technicality of the papers reviewed, it has been difficult in some cases to assess whether a work constitutes an "original task". Therefore, papers which were specifically mentioning the novelty of the proposed model, algorithm, or framework were considered to be original. The findings indicate a majority of 60% of the papers reportedly did an original work.

Human annotations. The study's second objective was to identify to which extent manually labeled data is being used in training datasets. Here, it is important to bear in mind that multiple datasets are usually being used to evaluate a single recommender system. When a study has mentioned *at least one* dataset which was annotated by humans, it was counted as using human annotations. Thus, it must not be interpreted that the proposed model uses *only* manually labeled datasets, but rather that it uses them to some extent. Interestingly, a vast majority of 86% work done in this domain is *not* making use of human-annotated data.

Instead, it has been observed that the main data sources use transactional data that has been publicly released by large vendors, such as MovieLens, Amazon, Last.FM, or Yelp, with more than one-third opting for MovieLens as part of their training and evaluation process. The exact proportions are shown in Table 2. When interpreting the table, it is important to note that the percentages indicate the number of papers that make use, but *are not limited to* that specific dataset. For example, 10% of the papers were using Yelp as part of their dataset choices, but it does not guarantee the exclusivity of other datasets.

The annotators. Another aim of the study was to further look at the annotators. Table 3 shows the population the annotators were drawn from. It is worth noting that the proportion is calculated from the papers which *actually reported* using some kind of annotations (14 in total). While there was no indication about the identity of the annotators in 28.57% of these papers, another 21.43% only mentioned they were crowdsourcing workers. Interestingly enough, no

Table 2
Most popular datasets used for training or evaluation.

Dataset	Count	Proportion
MovieLens	33	33%
Amazon	16	16%
Last.FM	10	10%
Yelp	10	10%

Table 3
Annotators population.

Annotators	Count	Proportion
Amazon Mechanical Turk	6	42.86%
Crowdsourcing workers	3	21.43%
The authors	1	7.14%
Not specified	4	28.57%

Table 4
Reported annotators number.

	Count	Proportion
Estimated number	0	0%
Actual number	9	64.29%

Table 5
Formal instructions, trainings, or pre-screening of annotators reported.

	Count	Proportion
Formal instructions	6	42.86%
Trainings	0	0%
Pre-screening	8	57.14%

work mentioned estimating the necessary number of annotators beforehand. However, the actual number of annotators was provided in most cases, as noted in Table 4.

Formal instructions, trainings, and pre-screening. A third objective was to identify if any type of pre-screening was done when selecting crowdsourcing workers, and whether formal instructions or trainings were provided beforehand. As listed in Table 5, this was rare, and training was never officially given.

Label quality. Further, the use of several metrics regarding label quality has been noted, such as multiple annotators labeling the same item, reported inter-annotator agreement, or label quality specification. Table 6 summarizes these results.

Link to datasets. Lastly, there was the observation regarding the datasets used, and more specifically, the extent to which the corresponding links are actually made available. The results

Table 6
Label quality metrics reported.

	Count	Proportion
Multiple annotators, same label	7	50%
Inter-annotator agreement	7	50%
Label quality	2	14.29%

Table 7
Link to dataset reported.

	Count	Proportion
All links provided	61	61%
Some links provided	31	31%
No link provided	8	8%

are summarized in Table 7.

4. Discussion

Drawing upon the findings mentioned in Section 3, the subsequent sections of this report will delve into three key aspects: the datasets used, the reproducibility of experiments, and the limitations inherent in this study. Through this discussion, the intent is to deepen the understanding of the significance and impact of the datasets on the overall study, while also highlighting areas for improvement and further investigation in the field.

4.1. Datasets Overview

Given the primary objective of this paper to enhance the understanding of current data collection and annotation practices, this section aims to explore the datasets employed by most of the papers by examining the key aspects with regard to their composition, quality, representativeness, and implications for the research outcomes. Given the relatively low percentage of human-annotated data actually being employed in the evaluation of recommender systems, a clear distinction of those datasets will be made in the exploration.

4.1.1. Human-annotated Data

When it comes to manually-annotated data, results reveal that 9 out of the 14 papers using these types of datasets mention crowdsourcing workers, mostly employed from Amazon Mechanical Turk (AMT). Furthermore, there are certain scenarios in which manual labor is necessary, such as evaluating the perceptions of explanations provided by recommendation systems that aim to offer explainability. The involvement of human annotators in this context contributes to the development of more effective and user-centric recommendation systems, and thus the main

purpose for them is to provide qualitative feedback. Below, a summary of the datasets that mentioned the use of human annotations is offered.

ReDial Dataset. ReDial comprises of dialogues in which users recommend movies to each other. Data is collected by pairing up AMT workers and giving them specific roles. Additional instructions are provided to improve data quality, such as using formal language and discussing at least four different movies per conversation. The collection is limited to English-speaking countries. Worker agreement on movie dialogue forms is used for validation [108].

TG-ReDial Dataset. TG-ReDial is a conversational dataset consisting of 129,392 utterances from 1,482 users. The data annotation process involves crowdsourcing workers from a specialized (unspecified) data annotation company. Each utterance is assigned to an annotator for labeling and an inspector for quality checking [109].

Beer Advocate Dataset. The dataset includes more than 1.5 million collected beer reviews spanning more than a decade until 2011. Ground truth labels were provided by external annotators, who annotated 1,000 reviews. While the inter-annotator agreement is reported, only 2 annotators were employed. It is noteworthy that the original dataset website indicates the data is no longer accessible, by BeerAdvocate's request [110].

CamRest676 Dataset. Human participants were recruited from AMT and assigned the roles of either a user or a wizard. The participants were instructed to compose conversations from the perspective of their assigned role. Users were given pre-specified goals to interact with the wizard, making the collected dialogue more representative of real-world scenarios. This approach aimed to ensure that the collected dialogue closely resembled actual user interactions [111].

Coat Shopping Dataset. The training data was generated by providing 270 AMT workers with a web shop interface. They were asked to find and rate their most desired coat from a selection of 300 items. Even though a link to trace the dataset was provided, in this case, special permissions are needed to actually access the data [112].

MyFitnessPal Dataset. To obtain this evaluation dataset, CrowdFlower was used to obtain human judgments of food substitutes. 100 food entries were randomly selected as target queries, and a ranked list of top-10 substitute candidates was generated for each query using two methods. CrowdFlower workers rated the suitability of 2,000 food substitute pairs on a 7-point Likert scale. Each pair was judged by three workers, and quality control was ensured using 57 ground truth questions. [113]

Our findings indicate that generally, some basic outlines regarding the annotation process are given. These include details such as the number of annotators, the instructions that they were given, or specifications regarding the quality of labels. When it comes to eligibility criteria, they mainly refer to proficiency in English, and no other complex requirements are specified. Contrary to the expectation, however, is the lack of information regarding the population these annotators were drawn from. For example, in the case of ReDial dataset, it is explicitly mentioned that the annotators reside in the US, Canada, UK, Australia, or New Zealand, but for other datasets this is not the case—while the representativeness of the annotator population may be worth considering, if a recommender system is intended to universally be effective.

To develop a full picture of the data, it is necessary to adopt a structured way of reporting the collection method, with extensive explanations of choices (e.g., why were these specific annotators chosen? What are the implications of employing these annotators from an ethical perspective?). When adopting more subjective criteria regarding the choice of annotators,

this introduces some degree of variability. Thus, by employing a more structured method for reporting, this would give the reader the possibility to make their own informed assessments.

4.1.2. Interaction Data

Since the majority of the reviewed papers leverage publicly available datasets, regarded as 'benchmarks' in the field, a discussion around the most popular ones follows.

MovieLens Dataset. More than 33% of the papers were using at least one version of this dataset [114], which contains ratings of movies. There are currently three benchmark versions of this dataset (10k, 1M, and 10M ratings), with the first two being employed by most papers. Interestingly enough, the data contained within these dates back to 1997-1998, and 2000, respectively. For that reason, its representativeness and relevance in terms of social aspects of nowadays' population could be debated. Furthermore, [115] explores biases and unfairness of this dataset in terms of two sensitive features, namely age, and gender. Their findings indicate that the biases are intrinsic to the dataset, regardless of the models used.

Yelp Dataset. This dataset contains data from Yelp, which is a review platform where users can leave reviews for businesses. It comprises approximately 7 million reviews given by almost 2 million users. Although the dataset is extensive, it still requires exploration to determine the population of users. [116] investigated the presence of biases in this dataset, mapping them to social, cultural, and political aspects.

While papers thoroughly justify the choice of algorithms, little to no explanations are given when it comes to the datasets used. A briefing consisting of the number of users and interactions is usually given, and the positioning of the choice relies on the fact that these datasets are widely used within the research domain. While the choice of these datasets could be motivated by wanting to compare novel models against a known benchmark, more emphasis should be put on shifting the focus from data *quantity* to data *quality*. A rather crucial question would be: To what extent are these data points representative of the population that the system aims to serve? Moreover, is it adequate for the specific domain of activity? Along the same line, [117] points out that there are no established metrics in place to determine the "goodness-of-data", as "goodness-of-fit" seems to be the preferred approach for most practitioners.

More in-depth exploration would be needed to reveal the quality or adequacy of the datasets employed by all reviewed papers. However, we argue that researchers should be more explicit regarding the rationale behind selecting a particular dataset, as the representativeness of datasets has been mentioned as a recurrent challenge of evaluating recommender systems [118].

Synthetic Data. Another interesting finding was the usage of synthetic datasets, especially in recommender systems that discussed bandits (i.e., recommender systems that are trying to balance the exploration phase of new items, with the exploitation phase of known items). It is therefore likely that the use of synthetic datasets comes from the need of training and evaluating on datasets that employ certain characteristics. While the generic outlines of these datasets are given, the findings indicate they are not usually being made publicly available.

Reproducibility of Experiments. Perhaps one of the most striking findings considers the extent to which experiments are actually reproducible. As reported in Table 8, 39% of the reviewed papers either provide no links to the datasets used or only provide *some* of the links (but not all), while this is relevant and even critical information for being able to reproduce a

work. Considering the widely recognized reputation of papers published in ACM, arguably, a rather standardized reporting practice should be deemed as necessary.

The absence of links to datasets was due to different situations: either the authors have made use of real-world datasets that are assumed to be well-known (and thus easy to trace), or they used synthetic/non-disclosable datasets. Regardless of the specific scenario, including the datasets represents a key part of a rigorous reporting procedure. By choosing not to do so, not only is the reproducibility of an experiment compromised but also the reliability of the results can then become questionable. In cases where there is really no possibility to disclose the datasets, a more comprehensive overview should be offered. While most of the time the numbers of users and interactions are given, the details could go beyond that. Does it consist of sensitive features? What is the population embedded in the dataset? And how representative it is for the domain in which the recommender system is employed? By doing so, it at least offers other researchers the relevant details to find a dataset with similar characteristics.

4.2. Study Limitations and Future Work

4.2.1. Limitations

Researcher bias. In systematic review studies, there may be risks of biases. Given the limited scope of the research, it might be that **sampling bias** has been introduced. As the literature on recommender systems contains thousands of papers, sampling only 100 of them might not be representative enough to draw generalizable conclusions. Furthermore, there may be **study design bias**. While established carefully and iteratively, there is no guarantee that the search criteria as mentioned in Section 2.2 are exhaustive. Thus, relevant papers may have been omitted from the study.

Time constraints. Reviewing and assessing literature on recommender systems can be time consuming, especially considering the complex, technical, and mathematical concepts discussed. As this research has been carried over a total period of 10 weeks with the lead author being an undergraduate student new to the field, more insights may still be obtainable from the collected data than were surfaced now.

Studies quality. One of the criteria employed to narrow down the search was the choice of a specific research database. Although it is one of the most appreciated within academia, there are certainly equally significant papers published in other journals. As noted by [119], the scientific contribution in itself does not necessarily rely on a journal's reputation, but on multiple indicators, amongst which actual influence in practical scenarios is noted. Hence, a greater focus on other types of metrics could be usefully explored in future research.

Results Interpretation. The interpretation of the results can be subject to limitations from two perspectives. Firstly, the limited amount of time, which did not allow for a more in-depth exploration of the datasets, and secondly, the subjective nature of the interpretations. It goes without saying that an expert in recommender systems might have judged the papers differently and might have based their conclusions on different evaluation criteria.

4.2.2. Further Work and Recommendations

Despite its limitations, this literature review is intended to at least serve as a starting point for a more extensive exploration of data collection and annotation practices within the domain of recommender systems. Further work is required to gain a more in-depth understanding of how these reporting practices are happening on a broader level, and what framework could possibly be adopted to include social factors in the discussion.

As for this, as previously mentioned, past work points out certain social and psychological factors that are inherent in the datasets, producing biases and ultimately, leading to skewed results. Hence, future research should aim to put more emphasis on societal impact from an interdisciplinary perspective. Furthermore, more attention is needed towards transparency on the data used to train the models. One way to tackle this issue would be to include a data sheet with specifications, similar to the one proposed by [120]. Examples of specifications include data composition, collection methods, data pre-processing, and intended use cases. To gain an in-depth understanding of each specification, several questions are posed. For instance, when it comes to data collection practices, [120] outline the need to understand different angles, such as sampling strategy, the timeframe of the collection, ethical review processes, or individuals involved in the collection process. In a similar fashion, [121] extensively introduces the so-called 'Data Cards' which aim to enhance transparency in the documentation process by offering a structured framework to work with. The outcome of the case studies conducted by the authors highlights crucial aspects that should be considered when working with a dataset, such as the problem space, intended or unsafe use cases, and data collection methods (including data sources and selection criteria). The adoption of such a structured framework allows one to derive deeper insights that might have otherwise been neglected, and can further assist in understanding the foundational components on which the ML model has been built on. Finally, whether it comes to annotated, synthetic, or interaction datasets, they should be linked and made available to ensure the reproducibility of experiments.

5. Conclusions

Recommender systems play a pivotal role in today's society, as they facilitate decision-making processes by helping users navigate extensive pools of information. It is known, however, that their ability to provide meaningful recommendations stems from training the recommendation model using large datasets. In this research paper, current data collection and annotation practices employed in scientific records were reviewed to identify whether techniques in state-of-the-art models take the quality of the data into account. The study examined several dimensions that influence data quality, including but not limited to the presence of human annotators, diversity within the annotator population, and label quality, whilst also looking at the public disclosure of datasets. One of the most significant findings to emerge from this analysis is that an overwhelming majority of practitioners employ just a few real-world benchmark datasets comprised of interaction data. Although standardized datasets are suitable to evaluate systems from an algorithmic perspective, arguably assessing the fit of the data is equally important to produce meaningful results. It is revealed that no robust reporting framework is in place and that often researchers fail to justify their dataset choices sufficiently. When it comes to annotated

data, general guidelines regarding the annotation process are usually given. However, it was found that little information is provided regarding the population from which the annotators were drawn. Consequently, a discussion was centered around the extent to which these datasets accurately represent the user population they aim to serve. Finally, the difficulty of reproducing experiments was uncovered, given the lack of links to datasets. Notwithstanding the relatively limited sample of the reviewed literature, this work offers valuable insights into the current state of training recommender system models and emphasizes the need for a consensus regarding rigorous reporting practices. Further research should be undertaken to explore how to establish an interdisciplinary framework to assess data quality and its fit for specific purposes when it comes to developing Machine Learning applications.

References

[1] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, J. Huang, Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 325–336. URL: https://dl.acm.org/doi/abs/10.1145/3351095.3372862.

[2] S. Kapania, A. S. Taylor, D. Wang, A hunt for the Snark: Annotator Diversity in Data Practices, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–15. URL: https://dl.acm.org/doi/abs/10.1145/3544548.3580645.

[3] M. Schrage, The recommender revolution, Technology review (2022).

[4] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, Ai & Society 35 (2020) 957–967. URL: https://link.springer.com/article/10.1007/s00146-020-00950-y.

[5] F. Ricci, L. Rokach, B. Shapira, Recommender systems: Techniques, applications, and challenges, Recommender Systems Handbook (2021) 1–35.

[6] A. Carrera-Rivera, W. Ochoa-Agurto, F. Larrinaga, G. Lasa, How-to conduct a systematic literature review: A quick guide for computer science research, MethodsX (2022) 101895.

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, International journal of surgery 88 (2021) 105906.

[8] J. Ormond, ACM Journals Shine in Latest Impact Factor Release, 2022. URL: https://www.acm.org/media-center/2022/august/acm-journals-impact-factor.

[9] F. Du, C. Plaisant, N. Spring, B. Shneiderman, Visual Interfaces for Recommendation Systems, ACM Transactions on Intelligent Systems and Technology 10 (2018) 1–23.

[10] Y. Bi, L. Song, M. Yao, Z. Wu, J. Wang, J. Xiao, DCDIR: A deep cross-domain recommendation system for cold start users in insurance domain, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.

[11] F. Lin, H.-P. Hsieh, A Joint Passenger Flow Inference and Path Recommender System for

Deploying New Routes and Stations of Mass Transit Transportation, ACM Transactions on Knowledge Discovery from Data 16 (2021) 1–36.
[12] W. Wang, H. Yin, Z. Huang, Q. Wang, X. Du, Q. V. H. Nguyen, Streaming Ranking Based Recommender Systems, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2018.
[13] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, Y. Zhang, Understanding echo chambers in e-commerce recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[14] P. Liu, L. Zhang, J. A. Gulla, Multilingual review-aware deep recommender system via aspect-based sentiment analysis, ACM Trans. Inf. Syst. 39 (2021) 1–33.
[15] G. de Souza Pereira Moreira, CHAMELEON, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, 2018.
[16] Y. Gu, Z. Ding, S. Wang, D. Yin, Hierarchical user profiling for e-commerce recommender systems, in: Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2020.
[17] Z. Xie, T. Yu, C. Zhao, S. Li, Comparison-based conversational recommender system with relative bandit feedback, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[18] H. Bharadhwaj, H. Park, B. Y. Lim, RecGAN, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, 2018.
[19] Y. Gu, Z. Ding, S. Wang, L. Zou, Y. Liu, D. Yin, Deep multifaceted transformers for multi-objective ranking in large-scale e-commerce recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2020.
[20] Z. Zhu, J. Kim, T. Nguyen, A. Fenton, J. Caverlee, Fairness among new items in cold start recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[21] V. Rubtsov, M. Kamenshchikov, I. Valyaev, V. Leksin, D. I. Ignatov, A hybrid two-stage recommender system for automatic playlist continuation, in: Proceedings of the ACM Recommender Systems Challenge 2018, ACM, 2018.
[22] J. Ma, Z. Zhao, X. Yi, J. Yang, M. Chen, J. Tang, L. Hong, E. H. Chi, Off-policy learning in two-stage recommender systems, in: Proceedings of The Web Conference 2020, ACM, New York, NY, USA, 2020.
[23] M. Chen, B. Chang, C. Xu, E. H. Chi, User response models to improve a REINFORCE recommender system, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2021.
[24] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, RippleNet, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018.
[25] T. Wu, E. K.-I. Chio, H.-T. Cheng, Y. Du, S. Rendle, D. Kuzmin, R. Agarwal, L. Zhang, J. Anderson, S. Singh, T. Chandra, E. H. Chi, W. Li, A. Kumar, X. Ma, A. Soares, N. Jindal, P. Cao, Zero-shot heterogeneous transfer learning from recommender systems to cold-

start search retrieval, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2020.
[26] S. Kang, J. Hwang, W. Kweon, H. Yu, Topology distillation for recommender system, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2021.
[27] H. Zhu, X. Li, P. Zhang, G. Li, J. He, H. Li, K. Gai, Learning Tree-based Deep Model for Recommender Systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018.
[28] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, T.-S. Chua, Estimation-action-reflection: Towards deep interaction between conversational and recommender systems, in: Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2020.
[29] H. Yang, T. Shen, S. Sanner, Bayesian critiquing with keyphrase activation vectors for VAE-based recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[30] M. Fang, G. Yang, N. Z. Gong, J. Liu, Poisoning Attacks to Graph-Based Recommender Systems, in: Proceedings of the 34th Annual Computer Security Applications Conference, ACM, 2018.
[31] H. Liu, X. Zhao, C. Wang, X. Liu, J. Tang, Automated embedding size search in deep recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[32] R. Zhan, K. Christakopoulou, Y. Le, J. Ooi, M. Mladenov, A. Beutel, C. Boutilier, E. Chi, M. Chen, Towards content provider aware recommender systems, in: Proceedings of the Web Conference 2021, ACM, New York, NY, USA, 2021.
[33] J. Tang, K. Wang, Ranking Distillation, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018.
[34] N. Liu, Y. Ge, L. Li, X. Hu, R. Chen, S.-H. Choi, Explainable recommender systems via resolving learning representations, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2020.
[35] X. Zhao, H. Liu, H. Liu, J. Tang, W. Guo, J. Shi, S. Wang, H. Gao, B. Long, AutoDim: Field-aware embedding dimension searchin recommender systems, in: Proceedings of the Web Conference 2021, ACM, New York, NY, USA, 2021.
[36] A. Ferraro, D. Bogdanov, J. Yoon, K. Kim, X. Serra, Automatic playlist continuation using a hybrid recommender system combining features from text and audio, in: Proceedings of the ACM Recommender Systems Challenge 2018, ACM, 2018.
[37] K. Luo, H. Yang, G. Wu, S. Sanner, Deep critiquing for VAE-based recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[38] C. Zhou, J. Ma, J. Zhang, J. Zhou, H. Yang, Contrastive learning for debiased candidate generation in large-scale recommender systems, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2021.
[39] X. Zhao, Q. Song, J. Caverlee, X. Hu, TrailMix, in: Proceedings of the ACM Recommender

Systems Challenge 2018, ACM, 2018.
[40] X. Ren, H. Yin, T. Chen, H. Wang, N. Q. V. Hung, Z. Huang, X. Zhang, Crsal, ACM Trans. Inf. Syst. 38 (2020) 1–40.
[41] M. Cheng, F. Yuan, Q. Liu, S. Ge, Z. Li, R. Yu, D. Lian, S. Yuan, E. Chen, Learning recommender systems with implicit feedback via soft target enhancement, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[42] Y. Li, M. Liu, J. Yin, C. Cui, X.-S. Xu, L. Nie, Routing Micro-videos via A Temporal Graph-guided Recommendation System, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019.
[43] Y. Sun, F. Yuan, M. Yang, G. Wei, Z. Zhao, D. Liu, A generic network compression framework for sequential recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[44] T. Huang, Y. Dong, M. Ding, Z. Yang, W. Feng, X. Wang, J. Tang, MixGCF, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2021.
[45] F. Lv, T. Jin, C. Yu, F. Sun, Q. Lin, K. Yang, W. Ng, SDM, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, ACM, 2019.
[46] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2020.
[47] S. Mu, Y. Li, W. X. Zhao, S. Li, J.-R. Wen, Knowledge-Guided Disentangled Representation Learning for Recommender Systems, ACM Transactions on Information Systems 40 (2021) 1–26.
[48] G. Wu, K. Luo, S. Sanner, H. Soh, Deep language-based critiquing for recommender systems, in: Proceedings of the 13th ACM Conference on Recommender Systems, ACM, 2019.
[49] Q. Tan, N. Liu, X. Zhao, H. Yang, J. Zhou, X. Hu, Learning to hash with graph neural networks for recommender systems, in: Proceedings of The Web Conference 2020, ACM, New York, NY, USA, 2020.
[50] W. Kweon, S. Kang, H. Yu, Bidirectional distillation for Top-K recommender system, in: Proceedings of the Web Conference 2021, ACM, New York, NY, USA, 2021.
[51] Q. Wu, H. Zhang, X. Gao, P. He, P. Weng, H. Gao, G. Chen, Dual Graph Attention Networks for Deep Latent Representation of Multifaceted Social Effects in Recommender Systems, in: The World Wide Web Conference, ACM, 2019.
[52] J. Zou, Y. Chen, E. Kanoulas, Towards question-based recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[53] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, Y. Zhang, Membership inference attacks against recommender systems, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, ACM, New York, NY, USA, 2021.
[54] H. Guo, J. Yu, Q. Liu, R. Tang, Y. Zhang, PAL, in: Proceedings of the 13th ACM Conference

on Recommender Systems, ACM, 2019.
[55] S. Kang, J. Hwang, W. Kweon, H. Yu, DE-RRD: A knowledge distillation framework for recommender system, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2020.
[56] I. Kostric, K. Balog, F. Radlinski, Soliciting user preferences in conversational recommender systems via usage-related questions, in: Fifteenth ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2021.
[57] D. Liu, C. Lin, Z. Zhang, Y. Xiao, H. Tong, Spiral of Silence in Recommender Systems, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019.
[58] W.-C. Kang, D. Z. Cheng, T. Chen, X. Yi, D. Lin, L. Hong, E. H. Chi, Learning multi-granular quantized embeddings for large-vocab categorical features in recommender systems, in: Companion Proceedings of the Web Conference 2020, ACM, New York, NY, USA, 2020.
[59] D. Antognini, B. Faltings, Fast multi-step critiquing for VAE-based recommender systems, in: Fifteenth ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2021.
[60] Y. Liu, K. Ge, X. Zhang, L. Lin, Real-time Attention Based Look-alike Model for Recommender System, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.
[61] D. Lian, H. Wang, Z. Liu, J. Lian, E. Chen, X. Xie, LightRec: A memory and search-efficient recommender system, in: Proceedings of The Web Conference 2020, ACM, New York, NY, USA, 2020.
[62] D. Kalimeris, S. Bhagat, S. Kalyanaraman, U. Weinsberg, Preference amplification in recommender systems, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2021.
[63] C. Qin, H. Zhu, C. Zhu, T. Xu, F. Zhuang, C. Ma, J. Zhang, H. Xiong, DuerQuiz, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.
[64] Y. Wang, D. Liang, L. Charlin, D. M. Blei, Causal inference for recommender systems, in: Fourteenth ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2020.
[65] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, X. Zhang, Fast-adapting and privacy-preserving federated recommender system, The VLDB Journal 31 (2021) 877–896.
[66] W. Sun, S. Khenissi, O. Nasraoui, P. Shafto, Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering, in: Companion Proceedings of The 2019 World Wide Web Conference, ACM, 2019.
[67] X. Xin, A. Karatzoglou, I. Arapakis, J. M. Jose, Self-Supervised Reinforcement Learning for Recommender Systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2020.
[68] C. Wu, D. Lian, Y. Ge, Z. Zhu, E. Chen, Triple adversarial learning for influence based poisoning attack in recommender systems, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2021.
[69] J. You, Y. Wang, A. Pal, P. Eksombatchai, C. Rosenburg, J. Leskovec, Hierarchical Temporal Convolutional Networks for Dynamic Recommender Systems, in: The World Wide Web

Conference, ACM, 2019.
[70] J. Wang, K. Ding, Z. Zhu, Y. Zhang, J. Caverlee, Key opinion leaders in recommendation systems, in: Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2020.
[71] W. Lin, X. Zhao, Y. Wang, T. Xu, X. Wu, AdaFS: Adaptive Feature Selection in Deep Recommender System, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, 2022.
[72] L. Zou, L. Xia, Z. Ding, J. Song, W. Liu, D. Yin, Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.
[73] G. Wang, Y. Zhang, Z. Fang, S. Wang, F. Zhang, D. Zhang, FairCharge, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (2020) 1–25.
[74] X. Xin, A. Karatzoglou, I. Arapakis, J. M. Jose, Supervised Advantage Actor-Critic for Recommender Systems, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, ACM, 2022.
[75] C. Wang, M. Zhang, W. Ma, Y. Liu, S. Ma, Modeling Item-Specific Temporal Dynamics of Repeat Consumption for Recommender Systems, in: The World Wide Web Conference, ACM, 2019.
[76] M. Wang, Y. Lin, G. Lin, K. Yang, X.-M. Wu, M2GRL: A multi-task multi-view graph representation learning framework for web-scale recommender systems, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2020.
[77] D. Zou, W. Wei, X.-L. Mao, Z. Wang, M. Qiu, F. Zhu, X. Cao, Multi-level Cross-view Contrastive Learning for Knowledge-aware Recommender System, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2022.
[78] H. Bharadhwaj, Explainable recommender system that maximizes exploration, in: Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion, ACM, New York, NY, USA, 2019.
[79] L. Yang, B. Liu, L. Lin, F. Xia, K. Chen, Q. Yang, Exploring clustering of bandits for online recommendation system, in: Fourteenth ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2020.
[80] T. Xiao, S. Wang, Towards Unbiased and Robust Causal Ranking for Recommender Systems, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, ACM, 2022.
[81] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, Personalized explanations for hybrid recommender systems, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, ACM, 2019.
[82] X. Zhao, Z. Zhu, Y. Zhang, J. Caverlee, Improving the estimation of tail ratings in recommender system with multi-latent representations, in: Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2020.
[83] Y. Zhou, K. Zhou, W. X. Zhao, C. Wang, P. Jiang, H. Hu, C^2-CRS, in: Proceedings of the

Fifteenth ACM International Conference on Web Search and Data Mining, ACM, 2022.
[84] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, Exploring High-Order User Preference on the Knowledge Graph for Recommender Systems, ACM Transactions on Information Systems 37 (2019) 1–26.
[85] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2020.
[86] W. L. Leite, S. Roy, N. Chakraborty, G. Michailidis, A. C. Huggins-Manley, S. D'Mello, M. K. S. Faradonbeh, E. Jensen, H. Kuang, Z. Jing, A novel video recommendation system for algebra: An effectiveness evaluation study, in: LAK22: 12th International Learning Analytics and Knowledge Conference, ACM, 2022.
[87] Y. Zheng, Utility-based multi-criteria recommender systems, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, ACM, 2019.
[88] Z. A. Pardos, W. Jiang, Designing for serendipity in a university course recommendation system, in: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, ACM, New York, NY, USA, 2020.
[89] N. R. Kermany, J. Yang, J. Wu, L. Pizzato, Fair-SRS: A Fair Session-based Recommendation System, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, ACM, 2022.
[90] M. Khwaja, M. Ferrer, J. O. Iglesias, A. A. Faisal, A. Matic, Aligning daily activities with personality, in: Proceedings of the 13th ACM Conference on Recommender Systems, ACM, 2019.
[91] K. Mahadik, Q. Wu, S. Li, A. Sabne, Fast distributed bandits for online recommendation systems, in: Proceedings of the 34th ACM International Conference on Supercomputing, ACM, New York, NY, USA, 2020.
[92] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, J.-R. Wen, Towards Universal Sequence Representation Learning for Recommender Systems, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, 2022.
[93] P. T. Nguyen, J. D. Rocco, D. D. Ruscio, L. Ochoa, T. Degueule, M. D. Penta, FOCUS: A Recommender System for Mining API Function Calls and Usage Patterns, in: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), IEEE, 2019.
[94] E. Shulman, L. Wolf, Meta decision trees for explainable recommendation systems, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, ACM, New York, NY, USA, 2020.
[95] Y. Wang, X. Zhao, T. Xu, X. Wu, AutoField: Automating Feature Selection in Deep Recommender Systems, in: Proceedings of the ACM Web Conference 2022, ACM, 2022.
[96] Q.-T. Truong, H. Lauw, Multimodal Review Generation for Recommender Systems, in: The World Wide Web Conference, ACM, 2019.
[97] L. V. Tran, Y. Tay, S. Zhang, G. Cong, X. Li, HyperML, in: Proceedings of the 13th International Conference on Web Search and Data Mining, ACM, 2020.
[98] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, Z. Wang, Knowledge-aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge

Discovery & Data Mining, ACM, 2019.
[99] A. H. Jadidinejad, C. Macdonald, I. Ounis, Using exploration to alleviate closed loop effects in recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2020.
[100] A. Kanakia, Z. Shen, D. Eide, K. Wang, A Scalable Hybrid Research Paper Recommender System for Microsoft Academic, in: The World Wide Web Conference, ACM, 2019.
[101] N. Yadav, A. K. Singh, Bi-directional Encoder Representation of Transformer model for Sequential Music Recommender System, in: Forum for Information Retrieval Evaluation, ACM, 2020.
[102] D. Tsumita, T. Takagi, Dialogue based recommender system that flexibly mixes utterances and recommendations, in: IEEE/WIC/ACM International Conference on Web Intelligence, ACM, 2019.
[103] C. Wu, D. Lian, Y. Ge, Z. Zhu, E. Chen, S. Yuan, Fight fire with fire: Towards robust recommender systems via adversarial poisoning training, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[104] D. Rafailidis, Bayesian Deep Learning with Trust and Distrust in Recommendation Systems, in: IEEE/WIC/ACM International Conference on Web Intelligence, ACM, 2019.
[105] J. Cho, S. Kang, D. Hyun, H. Yu, Unsupervised proxy selection for session-based recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[106] F. Eskandanian, N. Sonboli, B. Mobasher, Power of the Few, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, ACM, 2019.
[107] C. Lin, X. Liu, G. Xv, H. Li, Mitigating sentiment bias for recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021.
[108] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, 2018.
[109] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, 2020.
[110] J. McAuley, J. Leskovec, D. Jurafsky, Learning Attitudes and Attributes from Multi-aspect Reviews, in: 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 1020–1025.
[111] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, S. Young, A network-based end-to-end trainable task-oriented dialogue system, 2016.
[112] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, Recommendations as treatments: Debiasing learning and evaluation, 2016.
[113] P. Achananuparp, I. Weber, Extracting food substitutes from food diary via distributional similarity, 2016.
[114] F. M. Harper, J. A. Konstan, The MovieLens datasets: History and context, ACM transactions on interactive intelligent systems 5 (2016) 1–19. doi:10.1145/2827872.
[115] A. Gonzalez, F. Ortega, D. Perez-Lopez, S. Alonso, Bias and unfairness of collaborative filtering based recommender systems in MovieLens dataset, IEEE access: practical

innovations, open solutions 10 (2022) 68429–68439. URL: http://dx.doi.org/10.1109/access.2022.3186719.
[116] S. Choi, A. Pentland, An empirical study identifying bias in Yelp dataset, 2021.
[117] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, L. M. Aroyo, "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, 2021.
[118] E. Zangerle, C. Bauer, Evaluating Recommender systems: Survey and framework, volume=55, number=8, ACM computing surveys (2023) 1–38.
[119] A. Lindgreen, C. A. Di Benedetto, R. J. Brodie, Research quality: What it is, and how to achieve it, volume=99, issn=0019-8501, Industrial marketing management (2021) A13–A19.
[120] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, I. Daumé, Hal, K. Crawford, Datasheets for Datasets, 2018.
[121] M. Pushkarna, A. Zaldivar, O. Kjartansson, Data cards: Purposeful and transparent dataset documentation for responsible ai, 2022 ACM Conference on Fairness, Accountability, and Transparency (2022). URL: https://doi.org/10.1145/3531146.3533231. doi:10.1145/3531146.3533231.
[122] M. Burget, E. Bardone, M. Pedaste, Definitions and conceptual dimensions of Responsible Research and Innovation: A literature review, Science and engineering ethics 23 (2017) 1–19. URL: https://link.springer.com/article/10.1007/s11948-016-9782-1.
[123] J. Zhou, F. Chen, AI ethics: from principles to practice, AI society (2022). URL: https://link.springer.com/article/10.1007/s00146-022-01602-z.

A. Responsible Research

Responsible Research aims to unify conceptual dimensions such as anticipation, inclusion, responsiveness, and reflexivity, for the purpose of governing research and creating a positive societal impact. The emphasis is shifted from the *outcome* to the *actual process* of the research activity [122]. We discuss how these concepts were incorporated when conducting this research, covering integrity principles, reproducibility, and ethical aspects, to reflect more thoroughly on the societal implications of and ethical considerations behind our work.

Transparency and integrity. No financial support or funding has been given to conduct this research, and thus there is no conflict of interest arising from possible affiliations. Furthermore, in light of transparency, the limitations of this study have been extensively discussed in Section 4.2, taking into consideration possible biases, subjective criteria of results interpretation, study quality, and time constraints.

Reproducibility. The search criteria have been extensively explained in Section 2.2. However, it is important to note that the papers filtered are then selected based on the descending number of citations. As this number can potentially increase over time, there is no guarantee that replicating the same search string in the future will result in the exact same pool of papers as the one used when conducting this review. This is why we explicitly mentioned the date on which we conducted the search. Moreover, all of the data gathered during the review process

has been stored and made publicly available. To this extent, all analyzed papers, as well as their corresponding identifier, findings, or linked datasets are stored, so they can be further investigated if necessary. Thus, in case the search string might not be fully reusable in different settings, the reviewed papers can still be accessed in the future.

Ethical considerations. In light of Artificial Intelligence's growing popularity, supplementary efforts need to be made to establish clear guidelines with regard to AI ethics. To understand what is needed to make the ethical principles operable, [123] argue that AI ethics should be embedded in the whole AI lifecycle, starting with design, and following with data collection for training and testing purposes. Since recommender systems have been deeply integrated into our daily lives, having the ability to ultimately influence our decisions, it is crucial to address these kinds of considerations. By providing more clear insights into the current data annotation and collection practices observed throughout this research, the objective is to close the gap between Computer Science and other disciplines and encourage more interdisciplinary approaches.

B. CRediT author statement

Andra-Georgiana Sav: Methodology, Investigation, Data Curation, Writing – original draft; **Andrew M. Demetriou**: Conceptualization, Methodology, Resources, Writing – review & editing, Supervision; **Cynthia C. S. Liem**: Conceptualization, Writing – review & editing, Supervision, Project Administration.