Annotations Practices in Societally Impactful Machine Learning Applications

What are the recommender systems models actually trained on?

Andra-Georgiana Sav[1]

Supervisors: Cynthia Liem[1]**, Andrew Demetriou**[1]

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

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Andra-Georgiana Sav
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Andrew Demetriou, Frank Broz

Abstract

Machine Learning models are nowadays infused into all aspects of our lives. Perhaps one of its most common applications regards recommender systems, as they facilitate users' decision-making processes in various scenarios (e.g., e-commerce, social media, news, online learning, etc.). Training performed on large volumes of data is what ultimately drives such a system to provide meaningful recommendations, and yet there has been observed a lack of standardized practices when it comes to data collection and annotation methods for Machine Learning datasets. This research paper systematically identifies and synthesizes such processes by examining existing literature on recommender systems. The review includes 100 most-cited 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. A discussion is centered around the necessity of transitioning from reliance solely on algorithmic performance metrics in favor of prioritizing data quality and fit. Finally, valid 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.

1 Introduction

Automated systems are fueled by data –yet there has been observed a lack of standardized practices, processes or training of practitioners when it comes to annotating data for machine learning models, which consequently affects the reliability of the output produced. As also mentioned by Geiger et al., 2020 in their paper, most of the current Machine Learning research focuses on accuracy metrics to measure the correctness of the outputs, instead of also establishing qualitative data collection and annotation methods. Recent work of Kapania et al., 2023 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". The purpose of this paper is to generate a more in-depth understanding of the current practices in societally impactful Machine Learning applications, by conducting a systematic review of current literature.

As there might be discrepancies in findings relative to each research domain, the scope of this analysis will be narrowed down to a more specific area, namely recommender systems. Recommender systems have emerged as powerful instruments in nowadays' society, with use cases spanning across industries (e.g., media, banking, telecom, retail, etc.). As Schrage, 2022 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. Thus, it is important to adopt a multi-stakeholder perspective (Ricci et al., 2021) –from a consumer angle, these systems serve as a simplification of the search process, and prevent the information overload by only displaying relevant content (e.g., movies, videos, products, or even jobs recommendations). However, account providers' and system owners' interests might not be aligned with the users' original intents (e.g., increase sales of specific products, news propaganda). Therefore, it is important to understand the manner in which these systems provide recommendations and, ultimately, influence their users and shape their preferences.

In their literature review on recommender systems and the ethical challenges they pose, Milano et al., 2020 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. As users are unaware of how these systems actually work, one might be misled to believe that the recommendations meaningfully reflect their own interests. Moreover, exposure to only certain categories might incline their future choices towards those, thus reshaping the users' personal preferences. Thereby, having seen the data collection and annotation practices, it is clear that they play a pivotal role in understanding the underlying building blocks of these systems.

In this review, the following question is set to be answered: "*What are the recommender systems models actually trained on?*". This will be done by systematically capturing the extent to which the most cited 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, 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 (Section 2). Then, the findings section summarizes the outcomes of the review (Section 3). A discussion section provides an interpretation of the results focusing on the datasets and reports possible limitations and suggestions for further research (Section 4). Next, a section centered around responsible research follows (Section 5). Lastly, findings and key implications are briefly summarised in the paper's concluding chapter (Section 6).

2 Methodology

The current research paper is based on a systematic review method, as it 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" (Carrera-Rivera et al., 2022). This method has been initially used to gather relevant infor-

mation sources, but afterward, 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 (Page et al., 2021), 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, being placed in the 36th position out of 163 publications in Computer Science, Information Systems category (Ormond, 2022). The extensive list of reviewed papers can be seen in Table 9 in the Appendix.

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 the 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). For reproducibility purposes, the specific date on which the search string was run - May 8, 2023, is indicated. Table 1 presents the exact inclusion and exclusion criteria used to refine the scope of this search.

Table 1: Inclusion and exclusion criteria for research studies.

Criteria Type	Inclusion	Exclusion
Period	Publication year 2018-2023	Publication year prior to 2018
Source type	Research article, short paper, extended abstract	Survey, reviews, poster, invited talk, tutorial, work in progress, demonstration
Language	English	All other languages
Citation index	Top 100 most cited papers matching all criteria	
Search term in title only	recommender system(s), recommendation system(s)	
Search term in full text	supervised machine learning, supervised technique(s), supervised learning, supervised model, ground truth, gold standard	

The choice of reviewing the top 100 most cited papers is based on the need to narrow down the scope of the research, given the time constraint of 10 weeks. Furthermore, it is a good indicator of the current practices within the papers that create the most impact within this field.

2.3 Data collection process

The data collection has been entirely done by the author of this paper, which has 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 entailed. More specifically, inspired by the procedure adopted by Geiger et al., 2020, 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?
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 crowd-work 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, and double-checked by the reviewer. Results of a paper analysis are immediately noted down in the results table[1]. In case an answer to a sub-question is rather uncertain, this is subsequently 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. All of this information is accessible in the "Excluded Papers" table [2].

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.

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. Thus, as a rather tiny sample of papers from 2022 were actually analyzed, it could be argued that the findings might not necessarily be applicable to those.

[1]Table with collected data about reviewed papers:https://airtable.com/shrP0DCwzaMVdJRsA

[2]Table with papers excluded from the review process:https://airtable.com/shrbq6E0DxSo82rCo

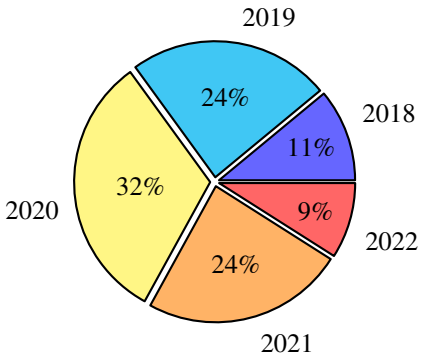

Figure 1: Reviewed papers distribution based on publication year.

Topic diversity. Given the suitability of certain techniques and approaches across various scenarios, the aim is to further classify the papers into recurrently encountered topic categories, as can be seen in Table 2.

Table 2: 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

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.

2.5 Risk of biases

In systematic review studies, there is a risk of biases arising, which has been assessed and further discussed in this section to provide full transparency.

Sample bias. Given the limited scope of the research, it might be the case 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.

Study design bias. The search criteria have been fully disclosed in Subsection 2.2. Although the process of selecting papers has been done iteratively, there is no guarantee that the list of relevant search terms is exhaustive. Thus, there exists a risk of having omitted relevant papers from the study.

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.

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 3.

Table 3: Most popular datasets used for training or evaluation.

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

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: What population were they drawn from? Was their number estimated beforehand? Or was the actual number specified?

Table 4 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, which were 14 in total. While there was no indication about the identity of the annotators in 28.57% of these papers, another 21.43% did only mention they were crowdsourcing workers.

It has been further studied to what extent is the number of annotators reported. Interestingly enough, no work mentioned estimating the necessary number of annotators. The actual number, however, was provided in most cases, as it can be noted in Table 5.

Table 4: Annotators population.

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

Table 5: Reported annotators number.

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

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 if formal instructions or trainings were provided beforehand. A summary of these results is reported in Table 6.

Table 6: Formal instructions, trainings, or pre-screening of annotators reported.

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

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 7 summarizes these results.

Table 7: Label quality metrics reported.

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

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 are summarized in Table 8.

Table 8: Link to dataset reported.

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

4 Discussion

Drawing upon the findings mentioned in Section 3, the subsequent sections of this report will delve into three key aspects, namely: the datasets used (Section 4.1), the reproducibility of experiments (Section 4.2), and the limitations inherent in this study (Section 4.3). 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 crowd-sourced workers, mostly employed from Amazon Mechanical Turk[3]. 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. (R. Li et al., 2018)

TG-ReDial Dataset. TG-ReDial is a conversational dataset consisting of 129,392 utterances from 1,482 users. The data annotation process involves crowd-sourced workers from a specialized data annotation company, which is not specified. Each utterance is assigned to an annotator for labeling and an inspector for quality checking. (K. Zhou, Zhou, et al., 2020)

Beer Advocate Dataset. Spanning more than a decade, the dataset includes more than 1.5 million collected reviews until 2011. The papers that utilize this dataset have incorporated ground truth labels provided by external annotators, which have annotated 1000 reviews. While the inter-annotator agreement is reported, there have been only 2 annotators employed. It is noteworthy that the original dataset website indicates that the data is no longer accessible, as per request of BeerAdvocate[4]. (McAuley et al., 2012)

CamRest676 Dataset. Human participants were recruited from Amazon Mechanical Turk and assigned the roles of either a user or a wizard. The participants were instructed to compose conversations from the perspective of their assigned

[3]https://www.mturk.com/

[4]https://www.beeradvocate.com/

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. (Wen et al., 2016)

Coat Shopping Dataset. The training data was generated by providing 270 Amazon Mechanical Turk 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. (Schnabel et al., 2016)

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. (Achananuparp and Weber, 2016)

Concluding Remarks. These findings provide some initial evidence that at least some basic outlines regarding the annotation process are generally 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, that is not the case. Given the small sample size of the datasets investigated, it does not suffice to draw a general conclusion. However, the representativeness of the annotator population is a concern worth considering, as it can raise questions regarding the quality of the annotated data. 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, it would at least 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 (Harper and Konstan, 2016), which contains ratings of movies. There are currently three benchmark versions of this dataset - with 10k, 1M, and 10M ratings, 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, Gonzalez et al., 2022 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, and thus reflect upon the presence of other sociological reasons.

Yelp Dataset[5]**.** 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. Choi and Pentland, 2021 investigated the presence of biases in this dataset, mapping them to social, cultural, or political aspects.

Concluding Remarks. While the choice of these datasets could be motivated by establishing performance comparisons of novel models, more emphasis should be put on shifting the focus from data *quantity* to data *quality*. The choice of the algorithms is thoroughly justified throughout the papers, however 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. 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, Sambasivan et al., 2021 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, it is argued that researchers should be more explicit regarding the rationale behind selecting a particular dataset, as one recurrent challenge of evaluating recommender systems regards exactly the representativeness of datasets (Zangerle and Bauer, 2023).

4.1.3 Synthetic Data

Another interesting finding was the usage of synthetic datasets, especially recurrent 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.

4.2 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

[5]https://www.yelp.com/dataset/

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 observed in one of the following cases: either the authors have made use of real-world datasets that are assumed to be well-known (and thus easy to trace), or they have 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.

4.3 Study Limitations and Future Work

4.3.1 Limitations

This section provides a critical assessment of the study and highlights its shortcomings with the intent to provide full transparency and encourage further exploration. To this extent, multiple perspectives are considered, such as biases, time constraints, studies quality, as well as results interpretation.

Bias Assessment. As also outlined in Section 2.5, there is a risk of biases arising when conducting systematic reviews. Although the search process has been carefully designed and performed iteratively, there is no guarantee that relevant search terms, and consequently relevant papers, were not excluded from this review, thus possibly introducing *sample and study design bias*.

Time constraints. Reviewing and assessing literature on recommender systems can be time-consuming, especially considering the complex, technical, and mathematical concepts discussed there. As this research has been carried over a total period of 10 weeks, it is rather difficult to derive more insights from the collected data.

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 Lindgreen et al., 2021, 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.

4.3.2 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 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 an interdisciplinary approach when designing Machine Learning applications. Furthermore, continued efforts are needed to adopt transparency when it comes to data used to train the models. One way to tackle this issue would be to include a "Data card" with data specifications, similar to the one proposed by Gebru et al., 2018 in their work. 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, Gebru outlines 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. Finally, whether it comes to annotated, synthetic, or interaction datasets, they should be linked and made available to ensure the reproducibility of experiments.

5 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 (Burget et al., 2017).

A discussion around how these concepts were incorporated when conducting this research follows, by closely examining integrity principles, ethical aspects, and reproducibility. The intention is to encourage the research community to reflect on the societal implications of their work, and openly address any ethical considerations.

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 4.3, 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 in-

crease 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. For that purpose, an indication regarding the specific date when the query was run is provided. 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, J. Zhou and Chen, 2022 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 a more multidisciplinary approach within this field.

6 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 a multidisciplinary framework to assess data quality and its fit for specific purposes when it comes to developing Machine Learning applications.

References

Achananuparp, P., & Weber, I. (2016). *Extracting food substitutes from food diary via distributional similarity*. http://arxiv.org/abs/1607.08807

Antognini, D., & Faltings, B. (2021). Fast multi-step critiquing for VAE-based recommender systems. *Fifteenth ACM Conference on Recommender Systems*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3460231.3474249

Bharadhwaj, H. (2019). Explainable recommender system that maximizes exploration. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3308557.3308699

Bharadhwaj, H., Park, H., & Lim, B. Y. (2018). RecGAN. *Proceedings of the 12th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3240323.3240383

Bi, Y., Song, L., Yao, M., Wu, Z., Wang, J., & Xiao, J. (2020). DCDIR: A deep cross-domain recommendation system for cold start users in insurance domain. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3397271.3401193

Burget, M., Bardone, E., & Pedaste, M. (2017). Definitions and conceptual dimensions of responsible research and innovation: A literature review. *Science and engineering ethics*, *23*(1), 1–19. https://link.springer.com/article/10.1007/s11948-016-9782-1

Carrera-Rivera, A., Ochoa-Agurto, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 101895. https://www.sciencedirect.com/science/article/pii/S2215016122002746/

Chen, M., Chang, B., Xu, C., & Chi, E. H. (2021). User response models to improve a REINFORCE recommender system. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3437963.3441764

Cheng, M., Yuan, F., Liu, Q., Ge, S., Li, Z., Yu, R., Lian, D., Yuan, S., & Chen, E. (2021). Learning recommender systems with implicit feedback via soft target enhancement. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-

org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462863

Cho, J., Kang, S., Hyun, D., & Yu, H. (2021). Unsupervised proxy selection for session-based recommender systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462958

Choi, S., & Pentland, A. (2021). An empirical study identifying bias in yelp dataset. https://dspace.mit.edu/bitstream/handle/1721.1/130685/1251779073-MIT.pdf?sequence=1

de Souza Pereira Moreira, G. (2018). CHAMELEON. *Proceedings of the 12th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3240323.3240331

Du, F., Plaisant, C., Spring, N., & Shneiderman, B. (2018). Visual interfaces for recommendation systems. *ACM Transactions on Intelligent Systems and Technology*, *10*(1), 1–23. https://doi.org/10.1145/3200490

Eskandanian, F., Sonboli, N., & Mobasher, B. (2019). Power of the few. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. https://doi.org/10.1145/3320435.3320464

Fang, M., Yang, G., Gong, N. Z., & Liu, J. (2018). Poisoning attacks to graph-based recommender systems. *Proceedings of the 34th Annual Computer Security Applications Conference*. https://doi.org/10.1145/3274694.3274706

Ferraro, A., Bogdanov, D., Yoon, J., Kim, K., & Serra, X. (2018). Automatic playlist continuation using a hybrid recommender system combining features from text and audio. *Proceedings of the ACM Recommender Systems Challenge 2018*. https://doi.org/10.1145/3267471.3267473

Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., & Zhang, Y. (2020). Understanding echo chambers in e-commerce recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl.acm.org/doi/10.1145/3397271.3401431

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, I., Hal, & Crawford, K. (2018). *Datasheets for datasets*. http://arxiv.org/abs/1803.09010

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

Gonzalez, A., Ortega, F., Perez-Lopez, D., & Alonso, S. (2022). Bias and unfairness of collaborative filtering based recommender systems in movielens dataset. *IEEE access: practical innovations, open solutions*, *10*, 68429–68439. http://dx.doi.org/10.1109/access.2022.3186719

Gu, Y., Ding, Z., Wang, S., & Yin, D. (2020). Hierarchical user profiling for e-commerce recommender systems. *Proceedings of the 13th International Conference on Web Search and Data Mining*. https://dl.acm.org/doi/abs/10.1145/3336191.3371827

Gu, Y., Ding, Z., Wang, S., Zou, L., Liu, Y., & Yin, D. (2020). Deep multifaceted transformers for multi-objective ranking in large-scale e-commerce recommender systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://dl.acm.org/doi/10.1145/3340531.3412697

Guo, H., Yu, J., Liu, Q., Tang, R., & Zhang, Y. (2019). PAL. *Proceedings of the 13th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3298689.3347033

Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM transactions on interactive intelligent systems*, *5*(4), 1–19. https://doi.org/10.1145/2827872

Hou, Y., Mu, S., Zhao, W. X., Li, Y., Ding, B., & Wen, J.-R. (2022). Towards universal sequence representation learning for recommender systems. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539381

Huang, T., Dong, Y., Ding, M., Yang, Z., Feng, W., Wang, X., & Tang, J. (2021). MixGCF. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3447548.3467408

Jadidinejad, A. H., Macdonald, C., & Ounis, I. (2020). Using exploration to alleviate closed loop effects in recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3397271.3401230

Kalimeris, D., Bhagat, S., Kalyanaraman, S., & Weinsberg, U. (2021). Preference amplification in recommender systems. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3447548.3467298

Kanakia, A., Shen, Z., Eide, D., & Wang, K. (2019). A scalable hybrid research paper recommender system for microsoft academic. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313700

Kang, S., Hwang, J., Kweon, W., & Yu, H. (2020). DE-RRD: A knowledge distillation framework for recommender system. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3340531.3412005

Kang, S., Hwang, J., Kweon, W., & Yu, H. (2021). Topology distillation for recommender system. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge*

Discovery & Data Mining. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3447548.3467319

Kang, W.-C., Cheng, D. Z., Chen, T., Yi, X., Lin, D., Hong, L., & Chi, E. H. (2020). Learning multi-granular quantized embeddings for large-vocab categorical features in recommender systems. *Companion Proceedings of the Web Conference 2020*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3366424.3383416

Kapania, S., Taylor, A. S., & Wang, D. (2023). A hunt for the snark: Annotator diversity in data practices. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. https://dl.acm.org/doi/abs/10.1145/3544548.3580645

Kermany, N. R., Yang, J., Wu, J., & Pizzato, L. (2022). FairSRS: A fair session-based recommendation system. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3488560.3502191

Khwaja, M., Ferrer, M., Iglesias, J. O., Faisal, A. A., & Matic, A. (2019). Aligning daily activities with personality. *Proceedings of the 13th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3298689.3347020

Kostric, I., Balog, K., & Radlinski, F. (2021). Soliciting user preferences in conversational recommender systems via usage-related questions. *Fifteenth ACM Conference on Recommender Systems*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3460231.3478861

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized explanations for hybrid recommender systems. *Proceedings of the 24th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3301275.3302306

Kweon, W., Kang, S., & Yu, H. (2021). Bidirectional distillation for Top-K recommender system. *Proceedings of the Web Conference 2021*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3442381.3449878

Lei, W., He, X., Miao, Y., Wu, Q., Hong, R., Kan, M.-Y., & Chua, T.-S. (2020). Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. *Proceedings of the 13th International Conference on Web Search and Data Mining*. https://dl.acm.org/doi/abs/10.1145/3336191.3371769

Leite, W. L., Roy, S., Chakraborty, N., Michailidis, G., Huggins-Manley, A. C., D'Mello, S., Faradonbeh, M. K. S., Jensen, E., Kuang, H., & Jing, Z. (2022). A novel video recommendation system for algebra: An effectiveness evaluation study. *LAK22: 12th International Learning Analytics and Knowledge Conference*. https://doi.org/10.1145/3506860.3506906

Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). *Towards deep conversational recommendations*. https://dl.acm.org/doi/epdf/10.5555/3327546.3327641

Li, Y., Liu, M., Yin, J., Cui, C., Xu, X.-S., & Nie, L. (2019). Routing micro-videos via a temporal graph-guided recommendation system. *Proceedings of the 27th ACM International Conference on Multimedia*. https://doi.org/10.1145/3343031.3350950

Lian, D., Wang, H., Liu, Z., Lian, J., Chen, E., & Xie, X. (2020). LightRec: A memory and search-efficient recommender system. *Proceedings of The Web Conference 2020*. https://dl.acm.org/doi/10.1145/3366423.3380151

Lin, C., Liu, X., Xv, G., & Li, H. (2021). Mitigating sentiment bias for recommender systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462943

Lin, F., & Hsieh, H.-P. (2021). 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*(1), 1–36. https://doi.org/10.1145/3451393

Lin, W., Zhao, X., Wang, Y., Xu, T., & Wu, X. (2022). AdaFS: Adaptive feature selection in deep recommender system. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539204

Lindgreen, A., Di Benedetto, C. A., & Brodie, R. J. (2021). Research quality: What it is, and how to achieve it. *Industrial marketing management*, *99*, A13–A19. https://www.sciencedirect.com/science/article/pii/S0019850121002121

Liu, D., Lin, C., Zhang, Z., Xiao, Y., & Tong, H. (2019). Spiral of silence in recommender systems. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3289600.3291003

Liu, H., Zhao, X., Wang, C., Liu, X., & Tang, J. (2020). Automated embedding size search in deep recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl.acm.org/doi/abs/10.1145/3397271.3401436

Liu, N., Ge, Y., Li, L., Hu, X., Chen, R., & Choi, S.-H. (2020). Explainable recommender systems via resolving learning representations. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3340531.3411919

Liu, P., Zhang, L., & Gulla, J. A. (2021). Multilingual review-aware deep recommender system via aspect-based sentiment analysis. *ACM Trans. Inf. Syst.*, *39*(2), 1–33. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3432049

Liu, Y., Ge, K., Zhang, X., & Lin, L. (2019). Real-time attention based look-alike model for recommender system. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3292500.3330707

Luo, K., Yang, H., Wu, G., & Sanner, S. (2020). Deep critiquing for VAE-based recommender systems. *Pro-*

ceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3397271.3401091

Lv, F., Jin, T., Yu, C., Sun, F., Lin, Q., Yang, K., & Ng, W. (2019). SDM. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. https://doi.org/10.1145/3357384.3357818

Ma, J., Zhao, Z., Yi, X., Yang, J., Chen, M., Tang, J., Hong, L., & Chi, E. H. (2020). Off-policy learning in two-stage recommender systems. *Proceedings of The Web Conference 2020*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3366423.3380130

Mahadik, K., Wu, Q., Li, S., & Sabne, A. (2020). Fast distributed bandits for online recommendation systems. *Proceedings of the 34th ACM International Conference on Supercomputing*. https://dl.acm.org/doi/10.1145/3392717.3392748

McAuley, J., Leskovec, J., & Jurafsky, D. (2012). Learning attitudes and attributes from multi-aspect reviews. *2012 IEEE 12th International Conference on Data Mining*, 1020–1025. https://ieeexplore.ieee.org/document/6413815

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

Mu, S., Li, Y., Zhao, W. X., Li, S., & Wen, J.-R. (2021). Knowledge-guided disentangled representation learning for recommender systems. *ACM Transactions on Information Systems*, *40*(1), 1–26. https://doi.org/10.1145/3464304

Nguyen, P. T., Rocco, J. D., Ruscio, D. D., Ochoa, L., Degueule, T., & Penta, M. D. (2019). FOCUS: A recommender system for mining API function calls and usage patterns. *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. https://doi.org/10.1109/icse.2019.00109

Ormond, J. (2022). Acm journals shine in latest impact factor release. https://www.acm.org/media-center/2022/august/acm-journals-impact-factor

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *International journal of surgery*, *88*, 105906. https://www.sciencedirect.com/science/article/pii/S1743919121000406

Pardos, Z. A., & Jiang, W. (2020). Designing for serendipity in a university course recommendation system. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3375462.3375524

Qin, C., Zhu, H., Zhu, C., Xu, T., Zhuang, F., Ma, C., Zhang, J., & Xiong, H. (2019). DuerQuiz. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3292500.3330706

Rafailidis, D. (2019). Bayesian deep learning with trust and distrust in recommendation systems. *IEEE/WIC/ACM International Conference on Web Intelligence*. https://doi.org/10.1145/3350546.3352496

Ren, X., Yin, H., Chen, T., Wang, H., Hung, N. Q. V., Huang, Z., & Zhang, X. (2020). CRSAL. *ACM Trans. Inf. Syst.*, *38*(4), 1–40. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3394592

Ricci, F., Rokach, L., & Shapira, B. (2021). Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook*, 1–35. https://link.springer.com/chapter/10.1007/978-1-0716-2197-4_1

Rubtsov, V., Kamenshchikov, M., Valyaev, I., Leksin, V., & Ignatov, D. I. (2018). A hybrid two-stage recommender system for automatic playlist continuation. *Proceedings of the ACM Recommender Systems Challenge 2018*. https://doi.org/10.1145/3267471.3267488

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://dl-acm-org.tudelft.idm.oclc.org/doi/abs/10.1145/3411764.3445518

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). *Recommendations as treatments: Debiasing learning and evaluation*. http://arxiv.org/abs/1602.05352

Schrage, M. (2022). The recommender revolution. *Technology review*. https://www.technologyreview.com/2022/04/27/1048517/the-recommender-revolution/

Shulman, E., & Wolf, L. (2020). Meta decision trees for explainable recommendation systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3375627.3375876

Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Wang, Z., & Zheng, K. (2020). Multi-modal knowledge graphs for recommender systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://dl.acm.org/doi/abs/10.1145/3340531.3411947

Sun, W., Khenissi, S., Nasraoui, O., & Shafto, P. (2019). Debiasing the human-recommender system feedback loop in collaborative filtering. *Companion Proceedings of The 2019 World Wide Web Conference*. https://doi.org/10.1145/3308560.3317303

Sun, Y., Yuan, F., Yang, M., Wei, G., Zhao, Z., & Liu, D. (2020). A generic network compression framework for sequential recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl.acm.org/doi/abs/10.1145/3397271.3401125

Tan, Q., Liu, N., Zhao, X., Yang, H., Zhou, J., & Hu, X. (2020). Learning to hash with graph neural networks for recommender systems. *Proceedings of The Web Conference 2020*. https://dl.acm.org/doi/abs/10.1145/3366423.3380266

Tang, J., & Wang, K. (2018). Ranking distillation. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3219819.3220021

Tran, L. V., Tay, Y., Zhang, S., Cong, G., & Li, X. (2020). HyperML. *Proceedings of the 13th International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3336191.3371850

Truong, Q.-T., & Lauw, H. (2019). Multimodal review generation for recommender systems. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313463

Tsumita, D., & Takagi, T. (2019). Dialogue based recommender system that flexibly mixes utterances and recommendations. *IEEE/WIC/ACM International Conference on Web Intelligence*. https://doi.org/10.1145/3350546.3352500

Wang, C., Zhang, M., Ma, W., Liu, Y., & Ma, S. (2019). Modeling item-specific temporal dynamics of repeat consumption for recommender systems. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313594

Wang, G., Zhang, Y., Fang, Z., Wang, S., Zhang, F., & Zhang, D. (2020). FairCharge. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *4*(1), 1–25. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3381003

Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., & Guo, M. (2018). RippleNet. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. https://doi.org/10.1145/3269206.3271739

Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., & Guo, M. (2019). Exploring high-order user preference on the knowledge graph for recommender systems. *ACM Transactions on Information Systems*, *37*(3), 1–26. https://doi.org/10.1145/3312738

Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., & Wang, Z. (2019). Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3292500.3330836

Wang, J., Ding, K., Zhu, Z., Zhang, Y., & Caverlee, J. (2020). Key opinion leaders in recommendation systems. *Proceedings of the 13th International Conference on Web Search and Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3336191.3371826

Wang, M., Lin, Y., Lin, G., Yang, K., & Wu, X.-M. (2020). M2GRL: A multi-task multi-view graph representation learning framework for web-scale recommender systems. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://dl.acm.org/doi/10.1145/3394486.3403284

Wang, Q., Yin, H., Chen, T., Yu, J., Zhou, A., & Zhang, X. (2021). Fast-adapting and privacy-preserving federated recommender system. *The VLDB Journal*, *31*(5), 877–896. https://doi.org/10.1007/s00778-021-00700-6

Wang, W., Yin, H., Huang, Z., Wang, Q., Du, X., & Nguyen, Q. V. H. (2018). Streaming ranking based recommender systems. *The 41st International ACM SIGIR Conference on Research &amp Development in Information Retrieval*. https://doi.org/10.1145/3209978.3210016

Wang, Y., Zhao, X., Xu, T., & Wu, X. (2022). AutoField: Automating feature selection in deep recommender systems. *Proceedings of the ACM Web Conference 2022*. https://doi.org/10.1145/3485447.3512071

Wang, Y., Liang, D., Charlin, L., & Blei, D. M. (2020). Causal inference for recommender systems. *Fourteenth ACM Conference on Recommender Systems*. https://dl.acm.org/doi/10.1145/3383313.3412225

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., & Young, S. (2016). *A network-based end-to-end trainable task-oriented dialogue system*. http://arxiv.org/abs/1604.04562

Wu, C., Lian, D., Ge, Y., Zhu, Z., & Chen, E. (2021). Triple adversarial learning for influence based poisoning attack in recommender systems. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3447548.3467335

Wu, C., Lian, D., Ge, Y., Zhu, Z., Chen, E., & Yuan, S. (2021). Fight fire with fire: Towards robust recommender systems via adversarial poisoning training. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462914

Wu, G., Luo, K., Sanner, S., & Soh, H. (2019). Deep language-based critiquing for recommender systems. *Proceedings of the 13th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3298689.3347009

Wu, Q., Zhang, H., Gao, X., He, P., Weng, P., Gao, H., & Chen, G. (2019). Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313442

Wu, T., Chio, E. K.-I., Cheng, H.-T., Du, Y., Rendle, S., Kuzmin, D., Agarwal, R., Zhang, L., Anderson, J., Singh, S., Chandra, T., Chi, E. H., Li, W., Kumar, A., Ma, X., Soares, A., Jindal, N., & Cao, P. (2020). Zero-shot heterogeneous transfer learning from recommender systems to cold-start search retrieval. *Proceedings of the 29th ACM International Conference on Information &amp Knowledge Manage-*

ment. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3340531.3412752

Xiao, T., & Wang, S. (2022). Towards unbiased and robust causal ranking for recommender systems. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3488560.3498521

Xie, Z., Yu, T., Zhao, C., & Li, S. (2021). Comparison-based conversational recommender system with relative bandit feedback. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462920

Xin, X., Karatzoglou, A., Arapakis, I., & Jose, J. M. (2020). Self-supervised reinforcement learning for recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3397271.3401147

Xin, X., Karatzoglou, A., Arapakis, I., & Jose, J. M. (2022). Supervised advantage actor-critic for recommender systems. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3488560.3498494

Yadav, N., & Singh, A. K. (2020). Bi-directional encoder representation of transformer model for sequential music recommender system. *Forum for Information Retrieval Evaluation*. https://doi.org/10.1145/3441501.3441503

Yang, H., Shen, T., & Sanner, S. (2021). Bayesian critiquing with keyphrase activation vectors for VAE-based recommender systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3463108

Yang, L., Liu, B., Lin, L., Xia, F., Chen, K., & Yang, Q. (2020). Exploring clustering of bandits for online recommendation system. *Fourteenth ACM Conference on Recommender Systems*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3383313.3412250

You, J., Wang, Y., Pal, A., Eksombatchai, P., Rosenburg, C., & Leskovec, J. (2019). Hierarchical temporal convolutional networks for dynamic recommender systems. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313747

Zangerle, E., & Bauer, C. (2023). Evaluating recommender systems: Survey and framework. *ACM computing surveys*, *55*(8), 1–38. http://dx.doi.org/10.1145/3556536

Zhan, R., Christakopoulou, K., Le, Y., Ooi, J., Mladenov, M., Beutel, A., Boutilier, C., Chi, E., & Chen, M. (2021). Towards content provider aware recommender systems. *Proceedings of the Web Conference 2021*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3442381.3449889

Zhang, M., Ren, Z., Wang, Z., Ren, P., Chen, Z., Hu, P., & Zhang, Y. (2021). Membership inference attacks against recommender systems. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3460120.3484770

Zhao, X., Liu, H., Liu, H., Tang, J., Guo, W., Shi, J., Wang, S., Gao, H., & Long, B. (2021). AutoDim: Field-aware embedding dimension searchin recommender systems. *Proceedings of the Web Conference 2021*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3442381.3450124

Zhao, X., Song, Q., Caverlee, J., & Hu, X. (2018). TrailMix. *Proceedings of the ACM Recommender Systems Challenge 2018*. https://doi.org/10.1145/3267471.3267479

Zhao, X., Zhu, Z., Zhang, Y., & Caverlee, J. (2020). Improving the estimation of tail ratings in recommender system with multi-latent representations. *Proceedings of the 13th International Conference on Web Search and Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3336191.3371810

Zheng, Y. (2019). Utility-based multi-criteria recommender systems. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. https://doi.org/10.1145/3297280.3297641

Zhou, C., Ma, J., Zhang, J., Zhou, J., & Yang, H. (2021). Contrastive learning for debiased candidate generation in large-scale recommender systems. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. https://dl-acm-org.tudelft.idm.oclc.org/doi/10.1145/3447548.3467102

Zhou, J., & Chen, F. (2022). Ai ethics: From principles to practice. *AI society*. https://link.springer.com/article/10.1007/s00146-022-01602-z

Zhou, K., Zhao, W. X., Bian, S., Zhou, Y., Wen, J.-R., & Yu, J. (2020). Improving conversational recommender systems via knowledge graph based semantic fusion. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://dl.acm.org/doi/abs/10.1145/3394486.3403143

Zhou, K., Zhou, Y., Zhao, W. X., Wang, X., & Wen, J.-R. (2020). *Towards topic-guided conversational recommender system*. http://arxiv.org/abs/2010.04125

Zhou, Y., Zhou, K., Zhao, W. X., Wang, C., Jiang, P., & Hu, H. (2022). $C^2 - CRS$. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3488560.3498514

Zhu, H., Li, X., Zhang, P., Li, G., He, J., Li, H., & Gai, K. (2018). Learning tree-based deep model for recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3219819.3219826

Zhu, Z., Kim, J., Nguyen, T., Fenton, A., & Caverlee, J. (2021). Fairness among new items in cold start recommender systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl-

acm-org.tudelft.idm.oclc.org/doi/10.1145/3404835.3462948

Zou, D., Wei, W., Mao, X.-L., Wang, Z., Qiu, M., Zhu, F., & Cao, X. (2022). Multi-level cross-view contrastive learning for knowledge-aware recommender system. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3477495.3532025

Zou, J., Chen, Y., & Kanoulas, E. (2020). Towards question-based recommender systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://dl.acm.org/doi/10.1145/3397271.3401180

Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., & Yin, D. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. https://doi.org/10.1145/3292500.3330668

A Appendix

Table 9: All reviewed papers.

All reviewed papers		
Du et al., 2018	Bi et al., 2020	F. Lin and Hsieh, 2021
W. Wang et al., 2018	Ge et al., 2020	P. Liu et al., 2021
de Souza Pereira Moreira, 2018	Gu, Ding, Wang, and Yin, 2020	Xie et al., 2021
Bharadhwaj et al., 2018	Gu, Ding, Wang, Zou, et al., 2020	Z. Zhu et al., 2021
Rubtsov et al., 2018	Ma et al., 2020	Chen et al., 2021
H. Wang et al., 2018	T. Wu et al., 2020	S. Kang et al., 2021
H. Zhu et al., 2018	Lei et al., 2020	H. Yang et al., 2021
Fang et al., 2018	H. Liu et al., 2020	Zhan et al., 2021
Tang and Wang, 2018	N. Liu et al., 2020	X. Zhao et al., 2021
Ferraro et al., 2018	Luo et al., 2020	C. Zhou et al., 2021
X. Zhao et al., 2018	Ren et al., 2020	Cheng et al., 2021
Y. Li et al., 2019	Y. Sun et al., 2020	Huang et al., 2021
Lv et al., 2019	R. Sun et al., 2020	Mu et al., 2021
G. Wu et al., 2019	Tan et al., 2020	Kweon et al., 2021
Q. Wu et al., 2019	J. Zou et al., 2020	Zhang et al., 2021
Guo et al., 2019	S. Kang et al., 2020	Kostric et al., 2021
D. Liu et al., 2019	W.-C. Kang et al., 2020	Antognini and Faltings, 2021
Y. Liu et al., 2019	Lian et al., 2020	Kalimeris et al., 2021
Qin et al., 2019	Y. Wang et al., 2020	Q. Wang et al., 2021
W. Sun et al., 2019	Xin et al., 2020	C. Wu, Lian, Ge, Zhu, and Chen, 2021
You et al., 2019	J. Wang et al., 2020	W. Lin et al., 2022
L. Zou et al., 2019	G. Wang et al., 2020	Xin et al., 2022
C. Wang et al., 2019	M. Wang et al., 2020	D. Zou et al., 2022
Bharadhwaj, 2019	L. Yang et al., 2020	Xiao and Wang, 2022
Kouki et al., 2019	X. Zhao et al., 2020	Y. Zhou et al., 2022
H. Wang, Zhang, Wang, et al., 2019	K. Zhou, Zhao, et al., 2020	Leite et al., 2022
Zheng, 2019	Pardos and Jiang, 2020	Kermany et al., 2022
Khwaja et al., 2019	Mahadik et al., 2020	Hou et al., 2022
Nguyen et al., 2019	Shulman and Wolf, 2020	Y. Wang et al., 2022
Truong and Lauw, 2019	Tran et al., 2020	
H. Wang, Zhang, Zhang, et al., 2019	Jadidinejad et al., 2020	
Kanakia et al., 2019	Yadav and Singh, 2020	
Tsumita and Takagi, 2019	C. Wu, Lian, Ge, Zhu, Chen, and Yuan, 2021	
Rafailidis, 2019	Cho et al., 2021	
Eskandanian et al., 2019	C. Lin et al., 2021	