ANNOTATION PRACTICES IN SOCIETALLY IMPACTFUL MACHINE LEARNING APPLICATIONS :

WHAT ARE THE RECOMMENDER SYSTEMS MODELS ACTUALLY TRAINED ON?

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 poster provides an **overview of these practices** and highlights the key findings from current literature.

1. BACKGROUND & MOTIVATION

There have been distinguished several ethical concerns regarding recommender systems, such as lack of user autonomy or personal identity. These can be addressed by increasing the transparency of user categorization or introducing more factual explanations.

To this extent, understanding the collection and annotation of data can play a pivotal role in understanding what the recommendations we receive are actually based on.

2. METHODOLOGY

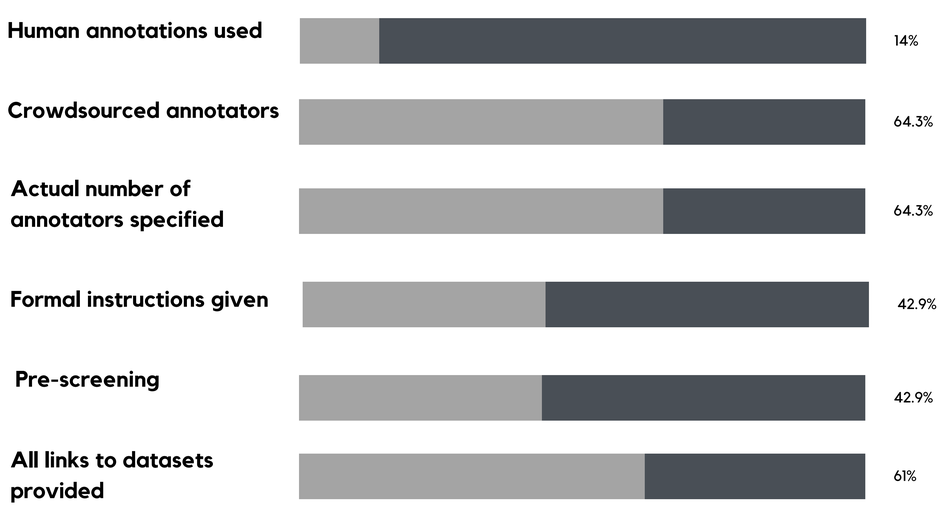
The research is carried out as a systematic review, drawing on the *top 100 most cited papers* from ACM Digital Library.

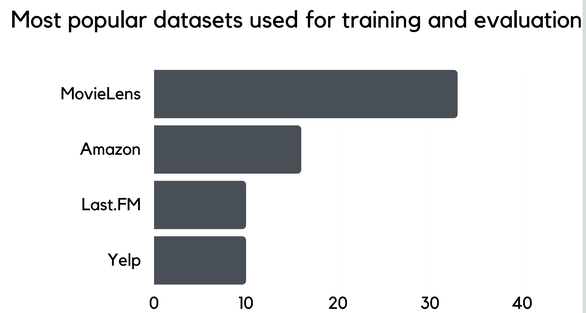
After the retrieval and review process was completed, further exploration of the datasets employed was done.

The papers reviewed adhere to the following inclusion criteria:

- **Publication year:** 2018-2023
- **Language:** English
- **Source type:** research article, short paper, extended abstract
- **Key terms in title:** recommender system(s), recommendation system(s)
- **Key terms in full text:** supervised machine learning, supervised technique(s), supervised learning, supervised model, ground truth, gold standard

MAIN FINDINGS





DATASET INSIGHTS

- **MOVIELENS.** The latest data was collected in 2000. Intrinsic biases were found with regard to sensitive features (i.e., age, gender).
- **YELP.** Interaction data comprising millions of reviews. Previous research studied the presence of biases from a social, political, and cultural perspective.
- **SYNTHETIC DATASETS.** Employed in some cases, but generally *not* disclosed.

3. TAKEAWAYS

- **Annotated datasets.** An overview of the annotation process is usually given, such as annotator number, label quality, and instructions given. However, explanations regarding the annotators' population are rarely given.
- **Benchmark datasets.** Benchmark datasets are preferred as to emphasize algorithmic performance metrics. However, little to no explanations are given regarding the *actual content* of a dataset, and its suitability for specific scenarios.
- **Reproducibility issues.** 39% of the reviewed papers provided insufficient links to datasets used for training or evaluation. The main assumption is that these datasets are well-known and widely used, or they are not publicly available.

4. RECOMMENDATIONS

- Adopt a **multidisciplinary approach** to overcome biases inherent in the datasets.
- Include a *"Data card"* with specifications of the datasets used (e.g., data composition, collection methods, ethical assessment, etc.).
- Implement **rigorous reporting practices** to ensure the reproducibility of experiments.

5. LIMITATIONS

- **Time constraints** of 10 weeks did not allow for a more in-depth exploration of the datasets
- **Sample and study design bias** might have occurred, as the search is not exhaustive.
- In terms of **papers' quality,** there might have been equally significant studies in other academic databases.
- **Interpretation of results** is limited to the technical understanding of the reviewer.

AFFILIATIONS

TU Delft
Delft University of Technology

AUTHOR
Andra-Georgiana Sav
E-mail: a.g.sav@student.tudelft.nl

PROFESSOR
Cynthia Liem

SUPERVISOR
Andrew Demetriou

REFERENCES

- 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? (pp. 325–336). https://doi.org/10.1145/3351095.3372862
- Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y
- Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets. ACM Transactions on Interactive Intelligent Systems, 5(4), 1–19. https://doi.org/10.1145/2827872
- https://www.yelp.com/dataset
- http://millionsongdataset.com/lastfm/
- https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews