Dataset quality within a societally impactful machine learning domain

An overview of data collection and annotation practices of the datasets used by papers published by the ACL

1. Background

• Datasets form a central part in the creation of a good machine learning model, however a lot of pitfalls to data selection[1] and annotations[2] exist
• Previous literature [3,4] suggests reporting practices for datasets used by papers across some domains is lackluster
• However, research regarding dataset formation practices is sparse, meaning more work within this domain would be valuable

2. Research questions

"What are the data collection and annotation practices of the datasets present in the most impactful papers of the ACL?"

1. which datasets are most often used by those papers? what is the overlap across different time periods?
2. how well do the most used datasets report on data collection and annotation practices? do they change for more recently impactful datasets?
3. how much information related to those practices is missing from those datasets? does this vary based on when those datasets were used?

3. Methodology

1. Selecting each of the top 25 most cited papers in the past 2, 5 and 15 years → see recent and general trends
2. Extracting the datasets used by each paper, and selecting for the top 20 by citation sum* for each timeframe
3. Going through each dataset paper, annotating it based on a schema previously used by [3,4], with some additional questions and analyzing the results

*citation sum: adding up the citations of all papers mentioning dataset

4. Results

• Table 1 shows the top 3 most used datasets
• Table 2 shows the amount of datasets per period, with 5 having the most datasets
• Table 3 shows the similarity between the periods in terms of datasets used: period 5 and 15 have a lot in common, while period 2 has very few in common with 5 and 15

Last 2 years

Dataset	Count
strategyqa	4
commonsenseqa	3
svamp	3

Last 5 years

Dataset	Count
glue	4
squad	4
multi-nli	3

Last 15 years

Dataset	Count
conll-2003	5
squad	4
cnn/dm	3

Overall

Dataset	Count
squad	9
conll-2003	7
glue	6

Table 1: Top 3 mentioned datasets by time frame

	2	5	15	overall
Avg datasets	3.478	7.05	4.25	4.821
Unique datasets	67	118	72	211
Total datasets	80	143	103	328

Table 2: Dataset statistics by time frame

	5 - 15	2 - 5	2 - 15
cosine similarity	0.501	0.116	0.067
common datasets	35.0	10.0	5.0

Table 3: Cosine similarity between periods

Research subquestion 1

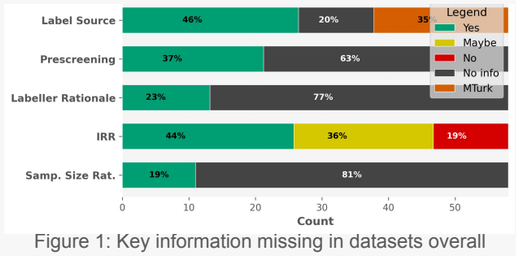

Figure 1: Key information missing in datasets overall

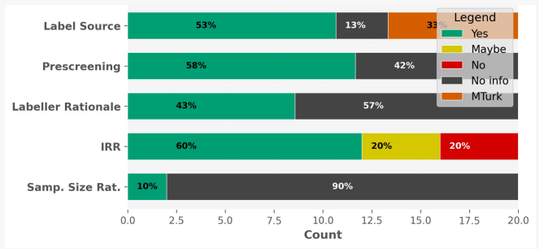

Figure 2: Key information missing in datasets past 2 years

Research subquestion 2

• Figure 1 shows the key information missing overall and figure 2 shows how it evolved with datasets used in the past 2 years
• The figures show a high (~⅓) use of MTurk (crowdsourcing platform) for label creation
• Prescreening, and Labeller Rationale are quite low, showing low consideration for selecting appropriate annotators for the task
• Both Prescreening and Labeller Rationale reporting grew in the past 2 years
• Inter-rater reliability (IRR) sometimes not calculated, lowering the credibility of the annotations - in past 2 years calculated more
• Item Sample Size Rationale is often not given, authors frequently "ending up" with this many items

Research subquestion 3

• Figure 3 shows the amount of information missing in the 3 timeframes and overall
• ⅓ of the information sought is missing overall
• there is less information missing from more recent periods, showing a trend towards more reporting
• however, ~¼ of the information sought is still missing from datasets used in the past 2 years

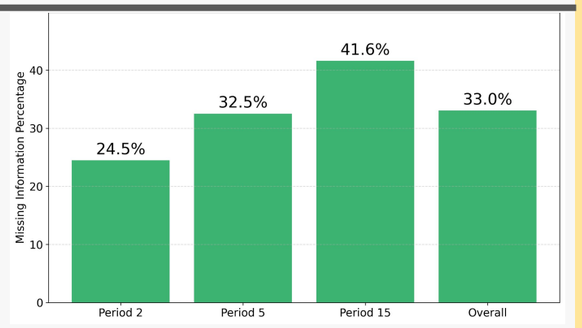

Figure 3: Total information missing per period and overall

5. Discussion

1. It can be seen that the recently used datasets are very different from datasets used in the past 5, 15 years, suggests (1) new and better datasets are used (2) domain changed.
2. The high use of MTurk as a label source may harm credibility [5]. Key aspects were identified where information is missing, but they are recently getting more documented.
3. ⅓ of the information is missing overall, with ¼ recently, but one should be cautious about the claims made by earlier work based on datasets of lower quality.

6. Limitations

- Time resources limited → no IRR calculated when annotating the datasets, as more emphasis was on annotating more datasets
- Although organizations implementing a checklist was suggested, this solution was not demonstrated to be effective

7. Conclusions

- Datasets used in the past 2 years have little overlap with the ones in the past 5, 15 years
- Reporting practices generally getting better, with less missing information, but information is still missing in some key areas
- In order to avoid such issues in the future, academic organizations should require their papers that use datasets to provide a dataset checklist, with items based on [6].

Authors

Alexandru Fazakas (A.Fazakas@student.tudelft.nl)
Responsible Professor: dr. Cynthia Liem
Supervisor: Andrew M. Demetriou

References

[1] J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, and A. Narayanan, "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES '22)*, Oxford, U.K., 2022, pp. 335–348, doi: 10.1145/3514094.3534196.
[2] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Mag.*, vol. 36, no. 1, pp. 15–24, Mar. 2015.
[3] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang, "'Garbage in, garbage out' revisited: What do machine learning application papers report about human-labeled training data?," *Quant. Sci. Stud.*, vol. 2, no. 3, pp. 795–827, 2021.
[4] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang, "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT* '20)*, ACM, 2020, pp. 325–336.
[5] H. Aguinis, I. Villamor, and R. S. Ramani, "MTurk research: Review and recommendations," *J. Manage.*, vol. 47, no. 4, pp. 823–837, 2020.
[6] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *CoRR*, vol. abs/1803.09010, 2018. [Online]. Available: http://arxiv.org/abs/1803.09010