Garbage In, Garbage Out?

Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang

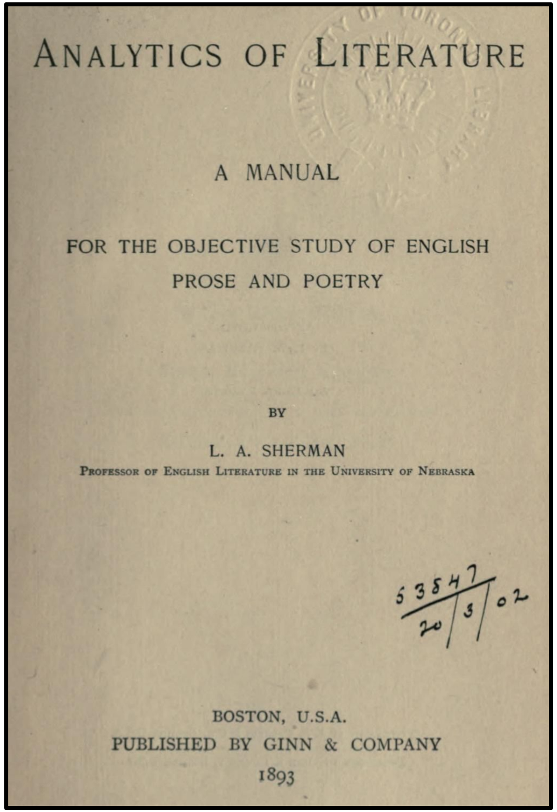
University of California, Berkeley
Berkeley Institute for Data Science (BIDS)
Division of Data Science & Information (DDSI)

This paper in four bullet points:

- Many of the ethical issues that arise in machine learning applications can be traced back to the quality of training data.
- The way training data is labeled by humans is often a form of structured content analysis, which has established best practices.
- RQ: How many papers in an application domain of ML --- classifiers trained on tweets --- report following these practices?
- A: It varies substantially, showing need for more focus on data labeling practices in ML education, evaluation, and regulation.

Structured content analysis (or closed coding)

An established method in the humanities and social sciences for generations.



ANALYTICS OF LITERATURE

A MANUAL

FOR THE OBJECTIVE STUDY OF ENGLISH PROSE AND POETRY

BY

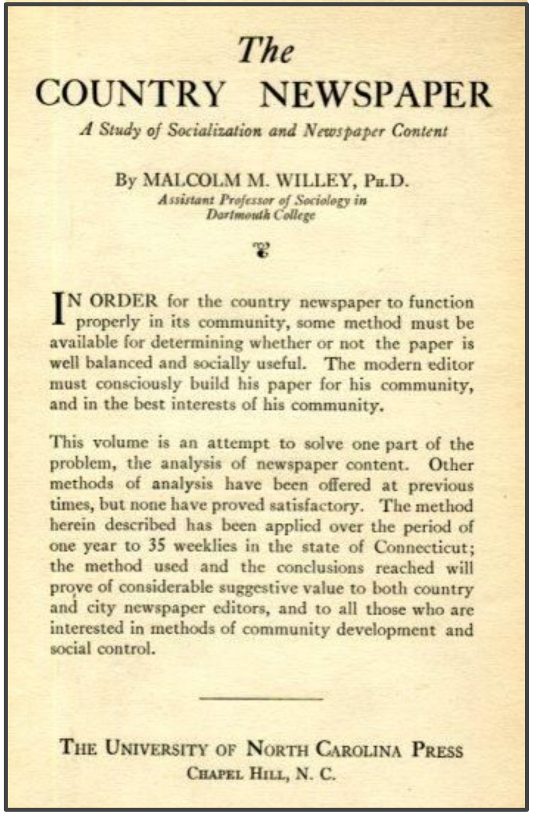
L. A. SHERMAN

PROFESSOR OF ENGLISH LITERATURE IN THE UNIVERSITY OF NEBRASKA

53847
20/3/02

BOSTON, U.S.A.
PUBLISHED BY GINN & COMPANY
1893

Sherman (1893)



The
COUNTRY NEWSPAPER
A Study of Socialization and Newspaper Content

By MALCOLM M. WILLEY, PH.D.
Assistant Professor of Sociology in Dartmouth College

IN ORDER for the country newspaper to function properly in its community, some method must be available for determining whether or not the paper is well balanced and socially useful. The modern editor must consciously build his paper for his community, and in the best interests of his community.

This volume is an attempt to solve one part of the problem, the analysis of newspaper content. Other methods of analysis have been offered at previous times, but none have proved satisfactory. The method herein described has been applied over the period of one year to 35 weeklies in the state of Connecticut; the method used and the conclusions reached will prove of considerable suggestive value to both country and city newspaper editors, and to all those who are interested in methods of community development and social control.

THE UNIVERSITY OF NORTH CAROLINA PRESS
CHAPEL HILL, N. C.

Willey (1923)

Structured content analysis best practices:

“a systematic and replicable method” (Riff, Lacy, and Frederick 2013)

1. Define a “coding scheme” with procedures, definitions, and examples.
2. Recruit and train multiple “coders” (or ”annotators”, “labelers”, or “reviewers”) with the coding scheme.
3. Have coders independently code at least a portion of the same items, then calculate “inter-annotator agreement” or “inter-rater reliability.”
4. Define and follow a process of “reconciliation” for disagreements, e.g. majority rule, talk to consensus, expert/leader decides.
5. Modify coding scheme, training, and/or reconciliation as needed.

Dataset of ML application papers trained on tweets

164 papers whose titles & abstracts matched searches for:

("machine learning", "classif", OR "supervi") AND ("twitter" OR "tweet")

135 randomly sampled from arxiv.org, 29 randomly sampled from Scopus

Published in 2010-2018

	Preprint never published	Postprint	Preprint of published paper	Non-ArXived (Scopus)	**Total**
Preprint never published	57	-	-	-	**57**
Refereed conference proceedings	-	40	17	23	**80**
Refereed journal article	-	8	7	6	**21**
Workshop paper	-	2	3	0	**5**
Dissertation	-	1	0	0	**1**
Total	**57**	**51**	**27**	**29**	164

Annotation process

Annotators: Five undergraduate students working for course credit independently reviewed each paper.

Reconciliation: Disagreement reconciled by talking to consensus, facilitated by the team leader, who made the final decision.

Iteration: Two rounds of annotation, after low IAA rates in round 1. Schema and instructions were modified after round 1.

IAA: mean percent total agreement across all questions was 84.4%.

Annota[redacted]

NO TIME FOR METHODS

Annotator[redacted] for course credit independently review[redacted]

Reconciliation: Disag[redacted]d by talking to consensus, facilitated by the tea[redacted]e the final decision.

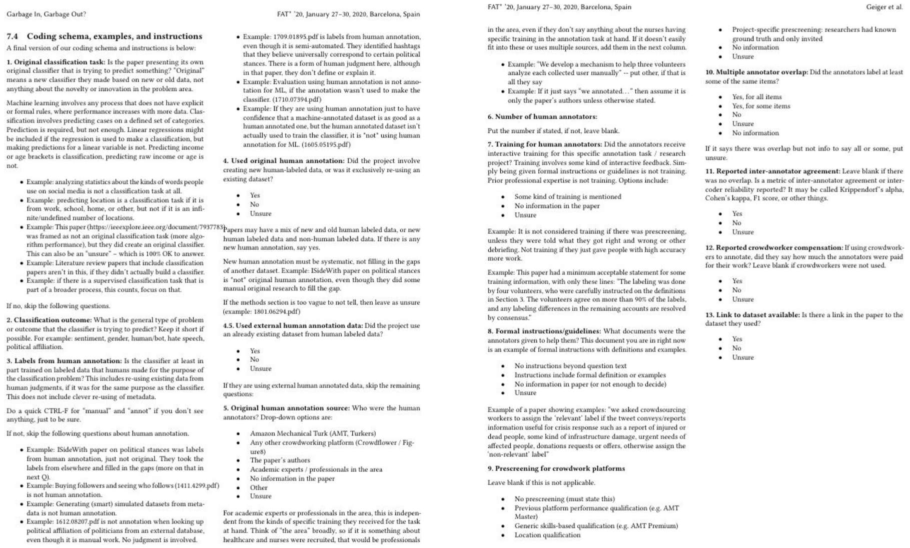
Iterati[redacted]s in round 1. Schem[redacted]

IAA: m[redacted]s was 84.4%.

Coding scheme

13 questions per item, which included definitions and examples for all items. Instructions, definitions, and examples totaled ~1,300 words.

Examples and definitions were iteratively updated in round 1 when borderline cases were discussed in reconciliation.



7.4 Coding schema, examples, and instructions

A final version of our coding schema and instructions is below:

1. Original classification task: Is the paper presenting its own original classifier that is trying to predict something? "Original" means a new classifier they made based on new or old data, not anything about the novelty or innovation in the problem area.

Machine learning involves any process that does not have explicit or formal rules, where performance increases with more data. Classification involves predicting cases on a defined set of categories. Prediction is required, but not enough. Linear regressions might be included if the regression is used to make a classification, but making predictions for a linear variable is not. Predicting income or age brackets is classification, predicting raw income or age is not.

- Example: analyzing statistics about the kinds of words people use on social media is not a classification task at all.
- Example: predicting location is a classification task if it is from work, school, home, or other, but not if it is an infinite/undefined number of locations.
- Example: This paper (https://ieeexplore.ieee.org/document/7937783) was framed as not an original classification task (more algorithm performance), but they did create an original classifier. This can also be an "unsure" – which is 100% OK to answer.
- Example: Literature review papers that include classification papers aren't in this, if they didn't actually build a classifier.
- Example: if there is a supervised classification task that is part of a broader process, this counts, focus on that.

If no, skip the following questions.

2. Classification outcome: What is the general type of problem or outcome that the classifier is trying to predict? Keep it short if possible. For example: sentiment, gender, human/bot, hate speech, political affiliation.

3. Labels from human annotation: Is the classifier at least in part trained on labeled data that humans made for the purpose of the classification problem? This includes re-using existing data from human judgments, if it was for the same purpose as the classifier. This does not include clever re-using of metadata.

Do a quick CTRL-F for "manual" and "annot" if you don't see anything, just to be sure.

If not, skip the following questions about human annotation.

- Example: ISideWith paper on political stances was labels from human annotation, just not original. They took the labels from elsewhere and filled in the gaps (more on that in next Q).
- Example: Buying followers and seeing who follows (1411.4299.pdf) is not human annotation.
- Example: Generating (smart) simulated datasets from metadata is not human annotation.
- Example: 1612.08207.pdf is not annotation when looking up political affiliation of politicians from an external database, even though it is manual work. No judgment is involved.
- Example: 1709.01895.pdf is labels from human annotation, even though it is semi-automated. They identified hashtags that they believe universally correspond to certain political stances. There is a form of human judgment here, although in that paper, they don't define or explain it.
- Example: Evaluation using human annotation is not annotation for ML, if the annotation wasn't used to make the classifier. (1710.07394.pdf)
- Example: If they are using human annotation just to have confidence that a machine-annotated dataset is as good as a human annotated one, but the human annotated dataset isn't actually used to train the classifier, it is "not" using human annotation for ML. (1605.05195.pdf)

4. Used original human annotation: Did the project involve creating new human-labeled data, or was it exclusively re-using an existing dataset?

- Yes
- No
- Unsure

Papers may have a mix of new and old human labeled data, or new human labeled data and non-human labeled data. If there is any new human annotation, say yes.

New human annotation must be systematic, not filling in the gaps of another dataset. Example: ISideWith paper on political stances is "not" original human annotation, even though they did some manual original research to fill the gap.

If the methods section is too vague to not tell, then leave as unsure (example: 1801.06294.pdf)

4.5. Used external human annotation data: Did the project use an already existing dataset from human labeled data?

- Yes
- No
- Unsure

If they are using external human annotated data, skip the remaining questions:

5. Original human annotation source: Who were the human annotators? Drop-down options are:

- Amazon Mechanical Turk (AMT, Turkers)
- Any other crowdworking platform (Crowdflower / Figure8)
- The paper's authors
- Academic experts / professionals in the area
- No information in the paper
- Other
- Unsure

For academic experts or professionals in the area, this is independent from the kinds of specific training they received for the task at hand. Think of "the area" broadly, so if it is something about healthcare and nurses were recruited, that would be professionals

in the area, even if they don't say anything about the nurses having specific training in the annotation task at hand. If it doesn't easily fit into these or uses multiple sources, add them in the next column.

- Example: "We develop a mechanism to help three volunteers analyze each collected user manually" -- put other, if that is all they say
- Example: If it just says "we annotated…" then assume it is only the paper's authors unless otherwise stated.

6. Number of human annotators:

Put the number if stated, if not, leave blank.

7. Training for human annotators: Did the annotators receive interactive training for this specific annotation task / research project? Training involves some kind of interactive feedback. Simply being given formal instructions or guidelines is not training. Prior professional expertise is not training. Options include:

- Some kind of training is mentioned
- No information in the paper
- Unsure

Example: It is not considered training if there was prescreening, unless they were told what they got right and wrong or other debriefing. Not training if they just gave people with high accuracy more work.

Example: This paper had a minimum acceptable statement for some training information, with only these lines: "The labeling was done by four volunteers, who were carefully instructed on the definitions in Section 3. The volunteers agree on more than 90% of the labels, and any labeling differences in the remaining accounts are resolved by consensus."

8. Formal instructions/guidelines: What documents were the annotators given to help them? This document you are in right now is an example of formal instructions with definitions and examples.

- No instructions beyond question text
- Instructions include formal definition or examples
- No information in paper (or not enough to decide)
- Unsure

Example of a paper showing examples: "we asked crowdsourcing workers to assign the 'relevant' label if the tweet conveys/reports information useful for crisis response such as a report of injured or dead people, some kind of infrastructure damage, urgent needs of affected people, donations requests or offers, otherwise assign the 'non-relevant' label"

9. Prescreening for crowdwork platforms

Leave blank if this is not applicable.

- No prescreening (must state this)
- Previous platform performance qualification (e.g. AMT Master)
- Generic skills-based qualification (e.g. AMT Premium)
- Location qualification
- Project-specific prescreening: researchers had known ground truth and only invited
- No information
- Unsure

10. Multiple annotator overlap: Did the annotators label at least some of the same items?

- Yes, for all items
- Yes, for some items
- No
- Unsure
- No information

If it says there was overlap but not info to say all or some, put unsure.

11. Reported inter-annotator agreement: Leave blank if there was no overlap. Is a metric of inter-annotator agreement or intercoder reliability reported? It may be called Krippendorf's alpha, Cohen's kappa, F1 score, or other things.

- Yes
- No
- Unsure

12. Reported crowdworker compensation: If using crowdworkers to annotate, did they say how much the annotators were paid for their work? Leave blank if crowdworkers were not used.

- Yes
- No
- Unsure

13. Link to dataset available: Is there a link in the paper to the dataset they used?

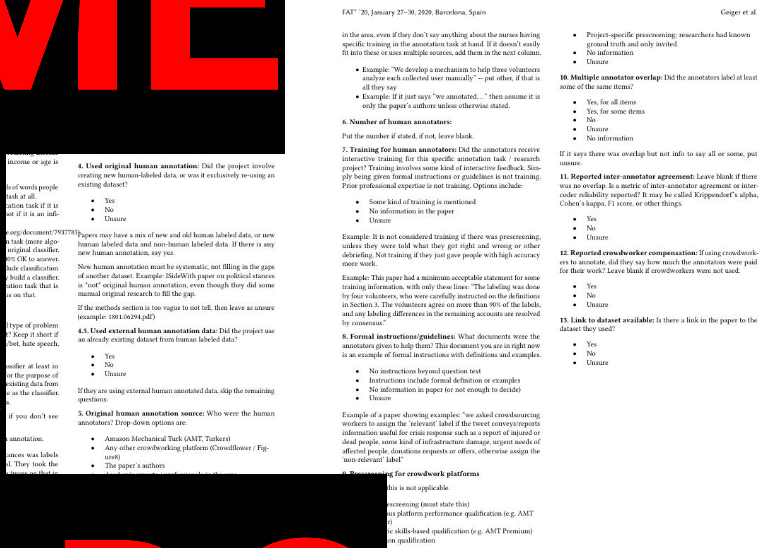
- Yes
- No
- Unsure

Coding

13 questions
included def
for all items.
definitions, and examp
~1,300 words.

Examples and definitio
iterativel
when bo
discusse

NO TIME FOR METHODS



FAT* '20, January 27–30, 2020, Barcelona, Spain — Geiger et al.

4. Used original human annotation: Did the project involve creating new human-labeled data, or was it exclusively re-using an existing dataset?

- Yes
- No
- Unsure

Papers may have a mix of new and old human labeled data, or new human labeled data and non-human labeled data. If there is any new human annotation, say yes.

New human annotation must be systematic, not filling in the gaps of another dataset. Example: ISideWith paper on political stances is *not* original human annotation, even though they did some manual original research to fill the gap.

If the methods section is too vague to not tell, then leave as unsure (example: 1801.06294.pdf)

4.5. Used external human annotation data: Did the project use an already existing dataset from human labeled data?

- Yes
- No
- Unsure

If they are using external human annotated data, skip the remaining questions:

5. Original human annotation source: Who were the human annotators? Drop-down options are:

- Amazon Mechanical Turk (AMT, Turkers)
- Any other crowdworking platform (Crowdflower / Figure8)
- The paper's authors

in the area, even if they don't say anything about the nurses having specific training in the annotation task at hand. If it doesn't easily fit into these or uses multiple sources, add them in the next column.

- Example: "We develop a mechanism to help three volunteers analyze each collected user manually" -- put other, if that is all they say
- Example: If it just says "we annotated…" then assume it is only the paper's authors unless otherwise stated.

6. Number of human annotators:

Put the number if stated, if not, leave blank.

7. Training for human annotators: Did the annotators receive interactive training for this specific annotation task / research project? Training involves some kind of interactive feedback. Simply being given formal instructions or guidelines is not training. Prior professional expertise is not training. Options include:

- Some kind of training is mentioned
- No information in the paper
- Unsure

Example: It is not considered training if there was prescreening, unless they were told what they got right and wrong or other debriefing. Not training if they just gave people with high accuracy more work.

Example: This paper had a minimum acceptable statement for some training information, with only these lines: "The labeling was done by four volunteers, who were carefully instructed on the definitions in Section 3. The volunteers agree on more than 90% of the labels, and any labeling differences in the remaining accounts are resolved by consensus."

8. Formal instructions/guidelines: What documents were the annotators given to help them? This document you are in right now is an example of formal instructions with definitions and examples.

- No instructions beyond question text
- Instructions include formal definition or examples
- No information in paper (or not enough to decide)
- Unsure

Example of a paper showing examples: "we asked crowdsourcing workers to assign the 'relevant' label if the tweet conveys/reports information useful for crisis response such as a report of injured or dead people, some kind of infrastructure damage, urgent needs of affected people, donations requests or offers, otherwise assign the 'non-relevant' label"

ng for crowdwork platforms

this is not applicable.

escreening (must state this)
us platform performance qualification (e.g. AMT
ic skills-based qualification (e.g. AMT Premium)
on qualification

- Project-specific prescreening: researchers had known ground truth and only invited
- No information
- Unsure

10. Multiple annotator overlap: Did the annotators label at least some of the same items?

- Yes, for all items
- Yes, for some items
- No
- Unsure
- No information

If it says there was overlap but not info to say all or some, put unsure.

11. Reported inter-annotator agreement: Leave blank if there was no overlap. Is a metric of inter-annotator agreement or intercoder reliability reported? It may be called Krippendorf's alpha, Cohen's kappa, F1 score, or other things.

- Yes
- No
- Unsure

12. Reported crowdworker compensation: If using crowdworkers to annotate, did they say how much the annotators were paid for their work? Leave blank if crowdworkers were not used.

- Yes
- No
- Unsure

13. Link to dataset available: Is there a link in the paper to the dataset they used?

- Yes
- No
- Unsure

Lots of borderline cases:

What is “machine learning,” anyway? Do simple linear regressions with a cutoff count? We said yes: any method without explicit rules where quality increases with the amount of data (Arthur Samuel’s definition)

What is “human labeling”? Does semi-automated labeling in bulk based on domain knowledge count (e.g. using #ProLife and #ProChoice to label political opinion)? We generally required discrete judgements on each item; hashtag example was external human annotation b/c the Twitter user “self-labeled” it.

What about using an automated method for labeling training data, but validating the classifier using individual human judgements? We said this isn’t human labeling.

For annotation source, who is an expert? We just looked for any claim of expertise beyond a member of the public, taking the authors’ at their word.

Lots of b[redacted]

NO TIME FOR METHODS

What is “machi[redacted]vith a cutoff count? We said yes: any m[redacted]vith the amount of data (Arthur Samuel[redacted]

What is “human labeling”? Do[redacted]ling in bulk based on domain knowledge count (e.g. using #[redacted] to label political opinion)? We generally required discrete ju[redacted]hashtag example was external human annotation b/c the Tw[redacted]

What abou[redacted]idating the classifier u[redacted]abeling.

For annota[redacted]expertise beyond a member[redacted]

Questions we asked:

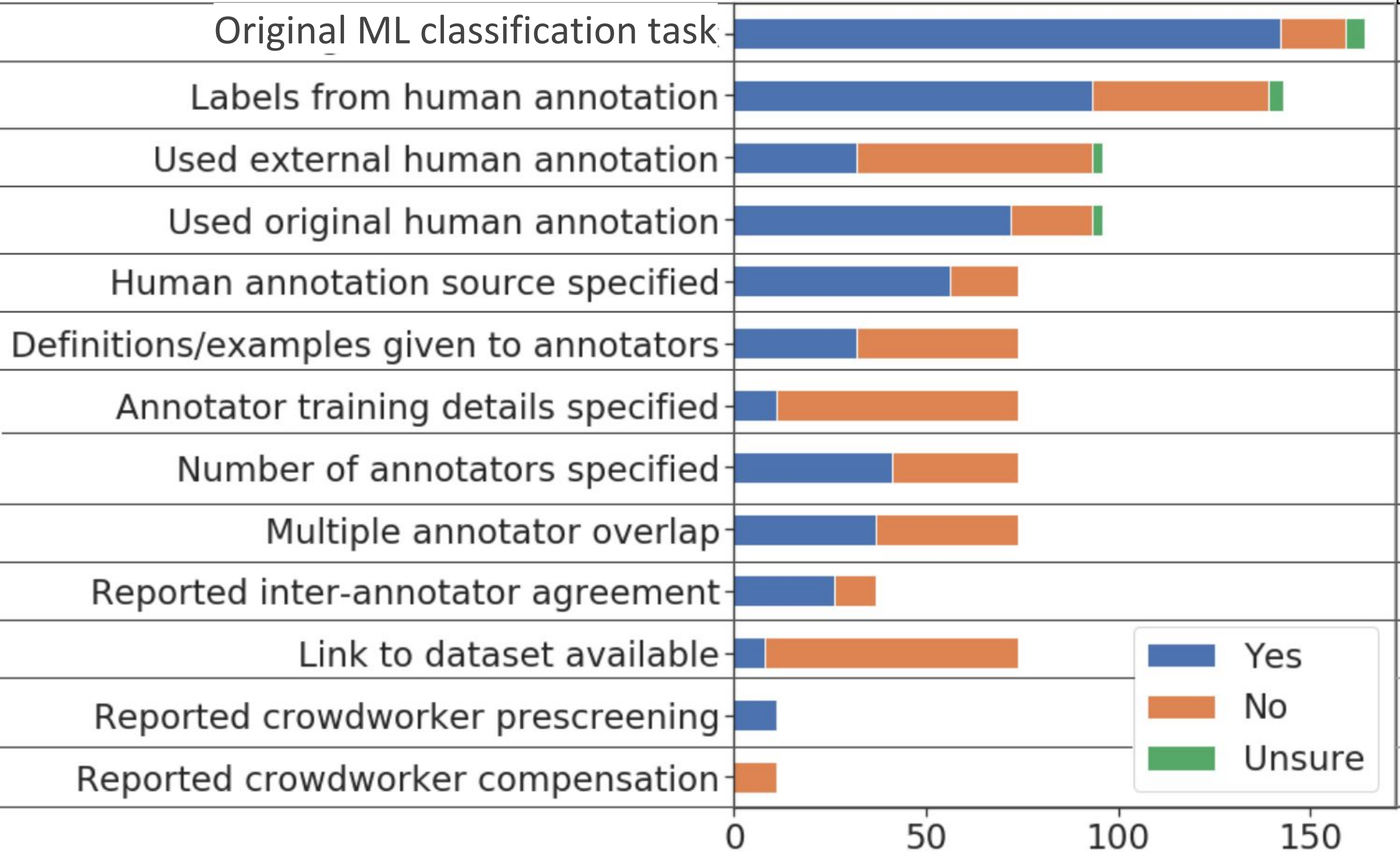
1. Is the paper presenting an original ML classification task?
2. Are the training data labels from human annotation?
3. Were the human labels from original labeling, an external dataset, or both?
4. Who labeled the dataset? (e.g. authors, turkers, experts)
5. Were the number of human annotators specified? (either total or per item)
6. Were instructions, formal definitions, or examples given to annotators?
7. Did annotators receive interactive training (beyond instructions/schema)?
8. For projects using crowdworkers, were annotators pre-screened?
9. Did multiple humans independently annotate every item (or some items)?
10. If so, were inter-annotator agreement metrics reported?
11. For projects using crowdworkers, was compensation reported?
12. Is there a link to the dataset available in the paper?

Question[redacted]

NO TIME FOR METHODS

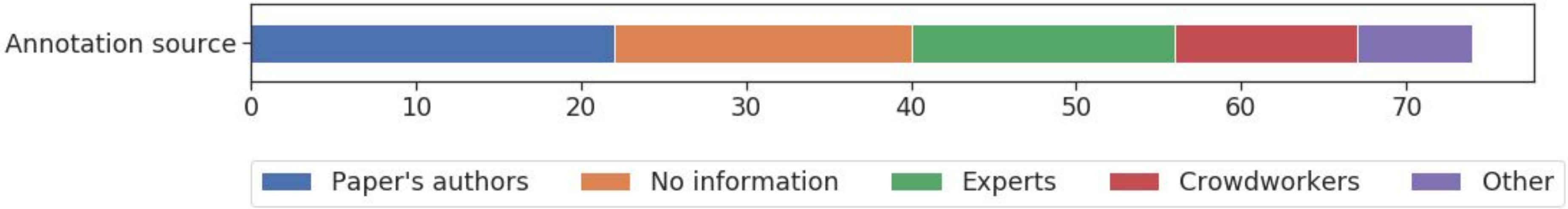
1. **Is the pap[redacted]k?**
2. **Are the tr[redacted]**
3. **Were the human label[redacted]g, an external dataset, or both?**
4. **Who labeled the datas[redacted]kers, experts)**
5. **Were the number of h[redacted]cified? (either total or per item)**
6. **Were instructions, for[redacted]amples given to annotators?**
7. **Did annotators receiv[redacted]beyond instructions/schema)?**
8. **For p[redacted]ened?**
9. **Did m[redacted]some items)?**
10. **If so,[redacted]**
11. **For p[redacted]ed?**
12. **Is the[redacted]**

Summary results



% yes
87%
65%
33%
75%
76%
43%
15%
55%
50%
70%
11%
100%
0%

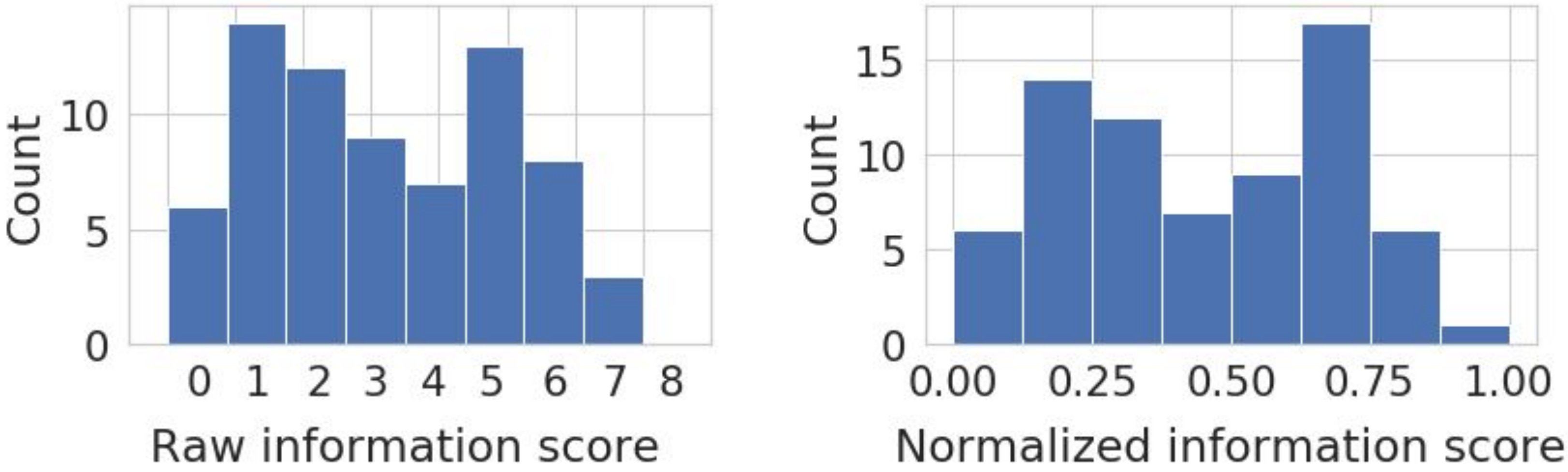
Human annotation source breakout:



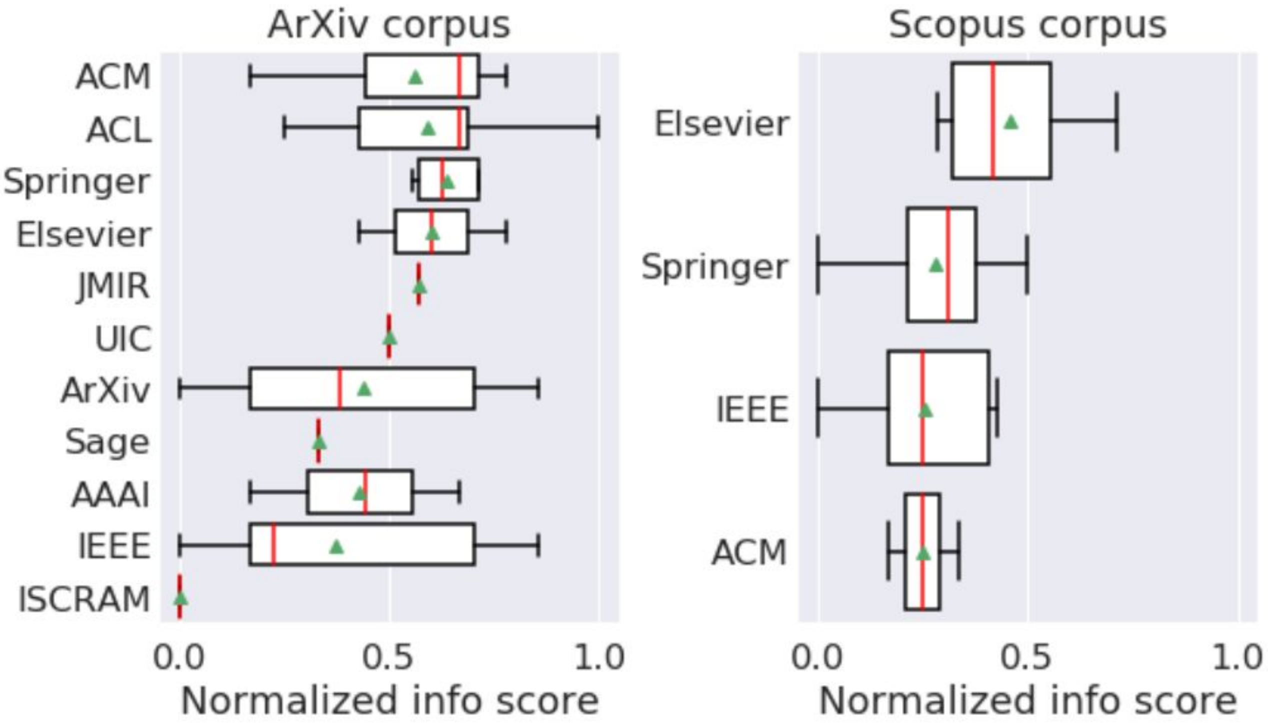
	Count	Proportion
Paper's authors	22	29.73%
No information	18	24.32%
Experts / professionals	16	21.62%
Amazon Mechanical Turk	3	4.05%
Other crowdwork	8	10.81%
Other	7	9.46%

Distribution of annotation information scores

A roughly bi-modial distribution suggests there are two populations of papers/studies.



Distribution of annotation information scores by publisher and corpus



Limitations and future work

We caution against over-generalizing these results! We have small sample sizes and arxiv.org is not representative.

Papers performing a classification task on Twitter data are also not representative, but do span many disciplines.

We are currently working on an expanded study, with additional questions, a refined process, and sampling from peer-reviewed literature across many application domains.

Discussion

Human annotation and labeling is as important as it is difficult.

We need to make space and time for methods and messiness!

Operationalization & construct validity decisions play out in the design of human annotation processes (see Jacobs et al, 2020). These should be made explicit!

Human annotation should be a core aspect of ML education and any structured transparency documentation process/regulation.

Discussion

For projects that presume a knowable & stable “ground truth”, scientific reproducibility is a classic principle:

> Is the labeling process described enough so any reader can, with sufficient resources, independently produce a substantively similar dataset?

Discussion

For projects that presume a knowable & stable "ground truth", scientific reproducibility is a classic principle:

> Is the labeling process described enough so any reader can, with sufficient resources, independently produce a substantively similar dataset?

What about when it is problematic to expect a "ground truth"? We can look to debates between quantitative/positivist social scientists and qualitative/interpretivist/critical social scientists and humanists (e.g. grounded theory); these are similar debates!

Thanks!

This work was funded in part by the Gordon & Betty Moore Foundation (Grant GBMF3834) and Alfred P. Sloan Foundation (Grant 2013-10-27).

This work was also supported by UC-Berkeley's Undergraduate Research Apprenticeship Program (URAP).

We thank many members of UC-Berkeley's Algorithmic Fairness & Opacity Group (AFOG) for providing invaluable feedback on this project!

Get in touch, especially if you have jobs/internships for some great undergrads!

-- R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang

BONUS SLIDES!

BERKELEY
Institute for
Data Science

GORDON AND BETTY
MOORE
FOUNDATION

ALFRED P. SLOAN
FOUNDATION

Software Tools and Environments

Reproducibility and Open Science

Career Paths and Alternative Metrics

Education and Training

Working Spaces and Culture

Data Science Studies

Other BIDS Data Science Studies research topics (aka shameless self-promotion)

- The *BIDS Best Practices in Data Science* Series
- Career paths and prospects in data science
- Sustainability of free and open-source software communities (see SciPy 2019 keynote)
- Reproducibility and open science
- Case studies of ML systems (e.g. Wikipedia's ORES)
- Integrating qualitative methods into data science
- The ArXiV Archive (arxiv.org metadata in tidy CSVs)
- The academic institutionalization of data science

Challenges of Doing Data-Intensive Research in Teams, Labs, and Groups: Report from the BIDS Best Practices in Data Science Series

R. Stuart Geiger[1†*], Dan Sholler[1,10†], Aaron Culich[3‡], Ciera Martinez[1,4‡], Fernando Hoces de la Guardia[5‡], François Lanusse[1,8,9‡], Kellie Ottoboni[1,2‡], Marla Stuart[1,6‡], Maryam Vareth[1,7‡], Nelle Varoquaux[1,2‡], Sara Stoudt[1,2‡], Stéfan van der Walt[1‡]

Best Practices for Fostering Diversity and Inclusion in Data Science

A Report from the Berkeley Institute for Data Science's *Best Practices in Data Science Series*

R. Stuart Geiger[1†*], Orianna DeMasi[1,10†], Aaron Culich[9‡], Andreas Zoglauer[1,3‡], Diya Das[1,4‡], Fernando Hoces de la Guardia[5‡], Kellie Ottoboni[1,2‡], Marsha Fenner[1‡], Nelle Varoquaux[1,2‡], Rebecca Barter[1,2‡], Richard Barnes[1,8‡], Sara Stoudt[1,2‡], Stacey Dorton[1‡], Stéfan van der Walt[1‡]

Best Practices for Managing Turnover in Data Science Groups, Teams, and Labs

A Report from the Berkeley Institute for Data Science's *Best Practices in Data Science Series*

Dan Sholler[1,2†*], Diya Das[1,3†], Fernando Hoces de la Guardia[4†], Chris Hoffman[5‡], François Lanusse[1,6,7‡], Nelle Varoquaux[1,11‡], Rolando Garcia[8‡], R. Stuart Geiger[1‡], Shana McDevitt[9‡], Scott Peterson[10‡], Sara Stoudt[1,11‡]

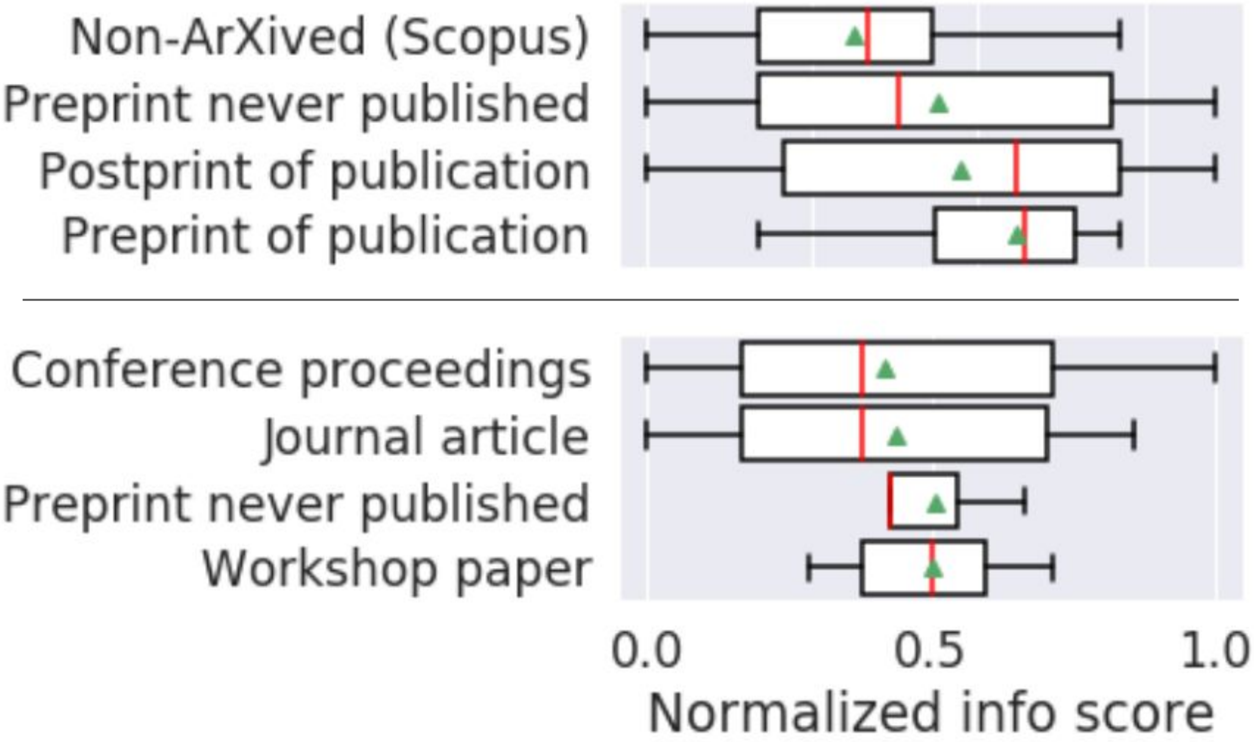
Resistance to Adoption of Best Practices

A Report from the Berkeley Institute for Data Science's *Best Practices in Data Science Series*

Dan Sholler[1,2†*], Sara Stoudt[1,3†], Chris Kennedy[1,4,5‡], Fernando Hoces de la Guardia[6‡], François Lanusse[1,7,8‡], Karthik Ram[1,2,9‡], Kellie Ottoboni[1,3‡], Marla Stuart[1,10‡], Maryam Vareth[1,11‡], Nelle Varoquaux[1,3‡], Rebecca Barter[1,3‡], R. Stuart Geiger[1‡], Scott Peterson[12‡], Stéfan van der Walt[1‡]

tinyurl.com/bidsbp

Distribution of annotation information scores by publication publication types



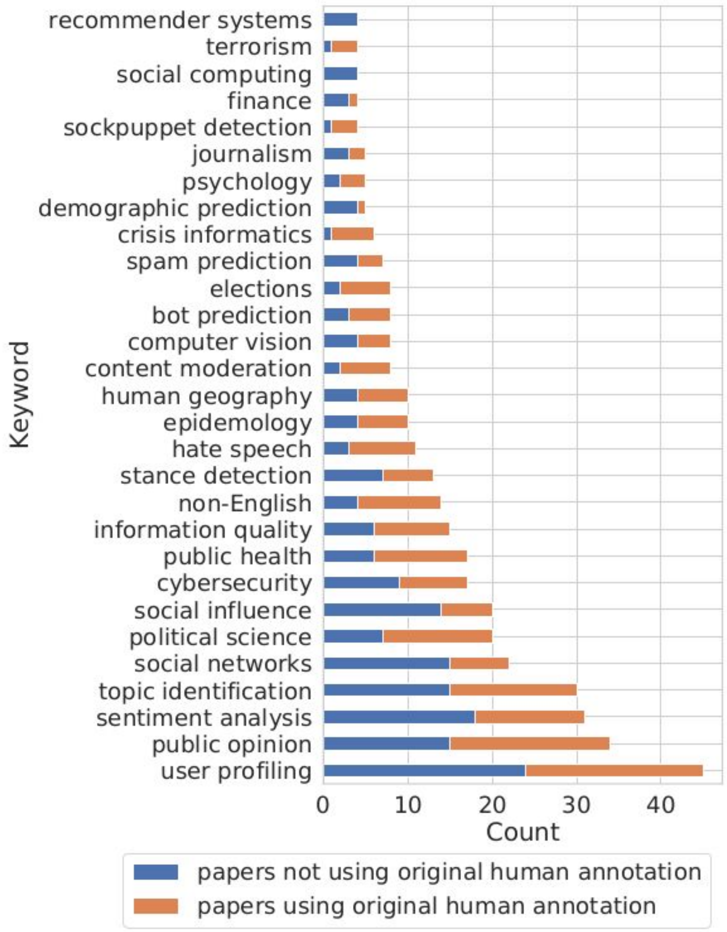


Figure 4: Plotting the distribution of papers by topical and disciplinary keywords, separated for papers using and not using original human annotation.

Year	# in ArXiv sample	# in Scopus sample
2010	1	0
2011	2	2
2012	2	2
2013	8	0
2014	5	4
2015	13	3
2016	29	5
2017	36	4
2018	39	9

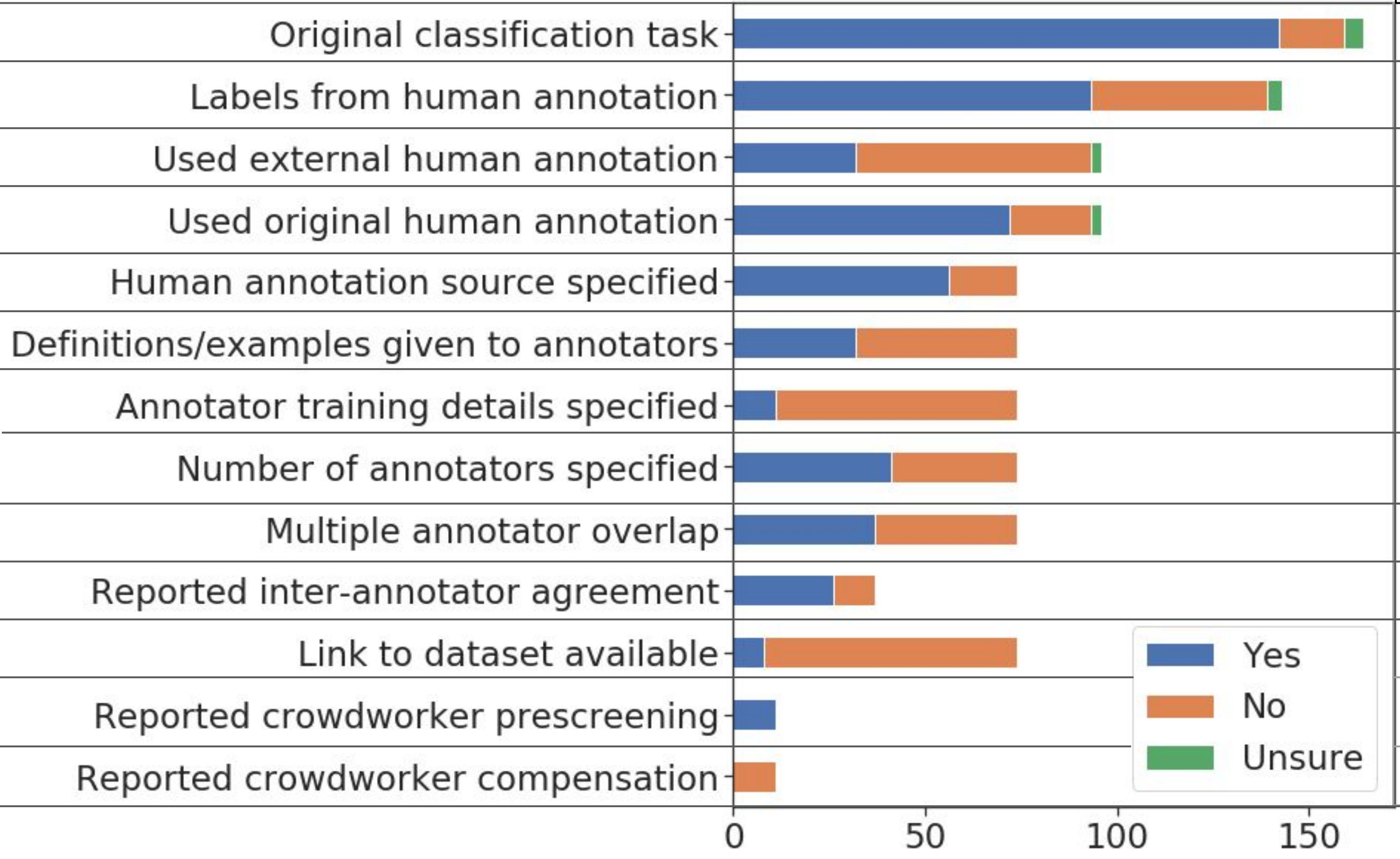
Table 14: Count of publications per year

From ArXiv sample		*From Scopus sample*	
Publisher	**Count**	**Publisher**	**Count**
ArXiv-only	58	Springer	7
ACM	20	ACM	5
IEEE	18	Elsevier	4
Springer	14	SPC	1
ACL	12		
Elsevier	4		
AAAI	3		
Sage	1		
CEUR	1		
PLoS	1		
UIC	1		
ISCRAM	1		
JMIR	1		

Table 15: Count of publishers from both samples

Question	% agreement, round 1	% agreement, round 2
original classification task	69.7%	93.9%
labels from human annotation	51.3%	82.9%
used original human annotation	72.0%	85.4%
used external human annotation	51.1%	63.4%
original human annotation source	44.3%	79.3%
number of annotators	38.2%	95.7%
training for human annotators	81.0%	84.8%
formal instructions	50.1%	82.9%
prescreening for crowdwork platforms	83.7%	89.0%
multiple annotator overlap	69.3%	81.7%
reported inter-annotator agreement	79.2%	83.5%
reported crowdworker compensation	94.9%	89.0%
link to dataset available	82.1%	86.0%
Mean score	66.7%	84.4%
Median score	69.5%	84.8%

Summary results



% yes
87%
65%
33%
75%
76%
43%
15%
55%
50%
70%
11%
100%
0%