Supplementary materials for *"Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?*

R. Stuart Geiger [1,†], **Dominique Cope** [2], **Jamie Ip** [2], **Marsha Lotosh** [3,†], **Aayush Shah** [2], **Jenny Weng** [2], **Rebekah Tang** [1,†]

***For correspondence:** stuart@stuartgeiger.com (RSG)

[†]The majority of the work on this paper was conducted when this author was affiliated with the University of California, Berkeley.

[1] University of California, San Diego
[2] University of California, Berkeley
[3] Webster Pacific

A. Corpus details

The tables below describe each corpus according to metadata provided by Scopus.

Appendix 0 Table 1. Publication year of papers by corpus

Year	Biomedical	Physical & Enviro Sci	Social Science	Grand Total
2013	2	3	4	9
2014	6	4	6	16
2015	6	7	5	18
2016	18	13	10	41
2017	14	15	13	42
2018	14	28	32	74
Grand Total	**60**	**70**	**70**	**200**

Appendix 0 Table 2. Classifier area/domain by corpus

	Biomedical	Physical & Enviro Sci	Social Science	Grand Total
Activities & actions	2	2	3	7
Biological	10	6	1	17
Demographic	0	2	3	5
Geo/ecological	1	9	3	13
Linguistic	2	3	19	24
Medical	25	12	6	43
Other	0	2	2	4
Physical	4	8	2	14
Soft/hardware	1	8	5	14
N/A	15	18	26	59
Grand Total	60	70	70	200

Appendix 0 Table 3. Publication type by corpus (note: all conference papers were peer reviewed according to Scopus)

	Biomedical	Physical & Enviro Sci	Social Science	Grand Total
Journal Article	50	40	31	121
Conference Paper	10	30	39	79
Grand Total	60	70	70	200

B. Labeling protocol and instructions

These are the instructions that were given to labelers. Note that we have removed links to specific papers that were used as examples, which can be provided upon request.

B.1. General purpose info

Put N/A if the question is not applicable. Some questions you put a - for no information.

"Unsure" is always an acceptable answer. It is OK to say that you are unsure. Some things are vague or complicated. Some of these papers are bad, wrong, messy, use words in completely different ways than is standard, shouldn't have been published at all, aren't actually about machine learning, and so on.

Flag complicated cases for discussion, use the last notes column. Some papers might have weird cases that require us to redefine our instructions, which is also OK. Don't spend too much time agonizing over small decisions.

"Coding" can mean manual annotation, this is a legacy from early 20th century linguistics.

If they have a link or citation to more info about their annotation, follow that. If it is an existing dataset from another project, that is out of scope. Don't try to search deeply for it if it isn't referenced in the paper.

If the paper is over 50 pages, skip it and flag for discussion.

If two rounds/stages of human annotation are involved, use the first one.

If there are multiple independent classifiers, some using human labels and others using machine labels, focus on the human-labeled classifier and ignore the machine-labeled one.

B.2. Questions

B.2.1. Original ML 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.

- Yes
- No
- Unsure / no information

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: if the paper is vague about if they actually built a classifier, choose unsure (link removed)

Example: any prediction on a linear/scalar value (not binned categories) is not classification (link removed)

Example: recommender systems are typically not classification (link removed)

Example: analyzing statistics about the kinds of words people use on social media is not a classification task

Example: predicting location is a classification task if it is from a set of locations (from work, school, home, or other) but not if it is an infinite/undefined number of locations. (link removed)

Example: This paper (link removed) 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.

B.2.2. Classification outcome:

What is the general type of problem or outcome that the classifier is trying to predict? This will be the label, typically. What is the end result, not how did they get there.

If multiple apply, put both and separate by comma

- Linguistic outcome
  - Sentiment, part of speech, stance, sarcasm, language, spam (if using NLP)
- Medical diagnosis
- Biological classification (non-medical)
- Physical classification
  - Astronomy, light, materials, chemicals, voltage
- Demographic classification
  - gender, race, class, political, job, transport
- Ecological/geographic/land use
  - Buildings
- Activities and actions
  - What are they doing? Includes gesture identification, tweeting while drinking, fraud/price manipulation
- Software and hardware entities

  - Includes various technologies that don't fit the above categories (esp not activities / actions)
  - Bot and other malware detection
- Other

B.2.3. Labels from human annotation:

Is the classifier at least in part trained on labeled data that involves human(s) who make individual judgments for specific items? This requires a human to make a judgement about individual cases.

- **Yes for all**
  - This is the typical case, where every item in the dataset used to train the classifier had a human make a judgment about the item
- **Yes for some**, which applies if *either*:
  - Human annotation was used to evaluate the classifier, but not train it; or,
  - There was some process of humans making judgments about some items, but then some kind of automated / mechanical way of scaling this to all items.
- **No / machine-labeled**
  - This includes fully-automated / machine-labeled ways of extracting labels, where a human is not involved in making each individual judgment
- **Implicit yes**
  - We know based on the subject matter that it had to be human labeled (e.g. patient medical data: 10.1109/CITSM.2017.8089245)
- **No information**
  - When we know for sure that there is no info in the paper and the context doesn't necessarily imply human labels (e.g. https://doi.org/10.1016/j.jtbi.2016.05.011)
- **Unsure**
  - This is when we are so confused by the paper that we don't even know if we can or can't answer it

If a human is told to do something that is recorded, where the recordings are the data and what they were told to do is the label, this counts as labels from human annotation. (e.g. (link removed) and (link removed))

This includes re-using existing data from human judgements, if it was for the same purpose as the classifier. This does not include clever re-using of metadata or human labeled data for a new purpose.

Setting a threshold for quantitative data is not human-labeled.

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

Example (link removed): this is yes for some, but very implicit. Unsupervised clustering used to re-label

Example: In medicine, if they were classifying for high blood pressure, and they use existing patient records with a cutoff of some number, this is not human annotation, it is "no / machine labeled". If they extract a physician's judgment from medical records, it is.

Example: labels were length of stay, not human annotation: (link removed)

> BUT: psych diagnostic tests that require judgement in scoring beyond skill are human judgement, although this is borderline.

Example: If the paper is using an external dataset that we know implicitly would have to use human annotation, but they don't say anything, put "Implicit yes." (link removed)

Example: 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 (link removed) is not human annotation.

Example: Generating (smart) simulated datasets from metadata is not human annotation.

Example: (link removed) not annotation when looking up political affiliation of politicians, even though it is manual work. No judgement is involved. "No."

Example: (link removed) identified hashtags that they believe universally correspond to certain political stances. This would be a kind of "self-annotation" by the tweet's author and therefore "yes for all"

Example: If they are using human annotation 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 counts as labels from human annotation --- see next question

Example: (link removed) did an inductive approach to finding accounts they thought were hate accounts, then used those tweets for training data. This is classification and human annotation.

Example: (link removed) -- they recruited patients from a parkinsons' clinic for evaluation, their classifier doesn't say what it was trained on. Put as "Yes, for some items"

On multi-stage processes: in general, focus on where humans are most in the loop with labeling. If there are multiple independent classifiers, some using human labels and others using machine labels, focus on the human-labeled classifier and ignore the machine-labeled one. E.g. (link removed) should be "yes for all" because we're just looking at the human-labeled one.

Unsure doesn't necessarily mean low quality: (link removed) -- eg cases where human judgement can be done at scale / semi-automated / rules

If not, skip the following questions about human annotation.

B.2.4. Used human annotation for training vs. evaluation:

If human annotation was used to generate labels, was it for the training dataset or just for evaluation?

- **No human annotation**
- **Human annotation for training data only**

  - This is the typical case, where labels are created then used to train the classifier. Often part of this data is held out as a test set to evaluate the classifier

- **Human annotation for evaluation only**
  - This is when they train the classifier using non human-labeled data, but use humans to either evaluate the validity of that dataset or the classifier.
- **Unsure**

B.2.5. Used original human annotation:

Did the project involve creating new human-labeled data (original), or was it exclusively re-using an existing dataset (external), or both? Think about this about as being organized by the researcher.

- **No original human annotation**
- **Only used original human annotation**
- **Only used external human annotation**
- **Used both original and external human annotation**
- **Unsure**

Papers may have a mix of new and old human labeled data, or new human labeled data and non-human labeled data.

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

For surveys, this counts as original if they ran the survey themselves, external if re-using someone else's survey data.

If the methods section is too vague to not tell, then leave as unsure (link removed)

If using transfer learning to augment a human-labeled classifier, this is original and external.

There may be overlap between the authors of a paper and a previous paper that presented the dataset. If a paper is treating a dataset/paper as a separate paper, consider it external, regardless of author overlap, even if they use language like "we previously." Example (link removed)

ONLY CONTINUE IF SOURCE INCLUDES ORIGINAL HUMAN ANNOTATION

Put N/A after for all items if not

B.2.6. Original human annotation source:

Who were the human annotators? Drop-down options are:

- **Amazon Mechanical Turk (AMT, Turkers)**
- **Other crowdworking platform (Crowdflower / Figure8)**
- **The paper's authors**
- **Other w/ claim of expertise**
- **Students (w/ no claim of expertise)**
- **Other (w/ no claim of expertise)**
- **Survey / self-reported data**

- **No information in the paper**
- **Unsure**

Survey / self reported includes self-annotation from social media (e.g. emoticons, hashtags), unless it was unstructured data that was then labeled by others. This also includes if a human is told to do something that is recorded, where the recordings are the data and what they were told to do is the label, this counts as labels from human annotation. (e.g. (link removed)

Do not consider data cleaning as part of annotation, unless it was systematic and involved looking at all or a random sample of the data.

"Other w/ claim of expertise" involves a specific claim that the annotators had qualifications beyond the average person. This is independent from the kinds of specific training they received for the task at hand.

Example: (link removed) psychiatrists made a determination of patients having or not having an eating disorder, based on surveys of patients. This is expert, not survey. But if they just had patients take one survey and the authors put them in different labels, this would be survey / self-reported.

Example: (link removed) they had volunteers collect data about road quality from smartphones, then they annotated their own data with labels. This is "survey / self-reported"

B.2.7. Prescreening for crowdwork platforms

Put N/A if this is not applicable, if they are not using crowdwork.

- **No prescreening (must state this)**
- **Previous platform performance qualification (e.g. AMT Master)**
- **Generic skills-based qualification (e.g. AMT Premium)**
- **Location qualification (country)**
- **Project-specific prescreening (e.g. inviting good crowdworkers back, doing their own prescreening)**
- **No information**
- **Unsure**

B.2.8. Annotator compensation:

Does the paper discuss how the annotators were compensated, if at all? If more than one applies, enter manually: list all and separate by commas

- **Money or gift cards**
- **Authorship on the paper**
- **Course credit**
- **Other compensation**
- **Volunteer / explicit no compensation**
- **No information**
- **Unsure**

If they are authors on the paper, put authorship

Example: if they said students were fluent in English and Hindi for a task involving sentiment analysis of English/Hindi tweets, this is a claim of expertise.

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.

B.2.9. 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. This is not about the qualifications of the annotator, but the training for this specific project.

Ex: (link removed) is some training data b/c "one drill session"

Options include:

- Some training detailed
- 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*."

B.2.10. 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

Ex (link removed): no instructions beyond question text

B.2.11. 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
- N/A

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

This is usually no for survey/self reported, because it often doesn't make sense. But could be, e.g. double checking if motion capture participation participants do the right motions.

B.2.12. Synthesis of annotator overlap (e.g. majority voting)

Put N/A if there was no overlap. There might be multiple means stated, pick the primary synthesis method.

- Qualitative / discussion
  - Process should involve some interaction between labelers
- Quantitative / no discussion
  - Includes majority vote, Bayesian, recruit a tiebreaker
- Other
- No information
- Unsure
- N/A -- if there was no multiple overlap

Use the default synthesis mechanism. For example: "Labelers met in person to discuss cases of disagreement. Where consensus could not be reached, the project lead made the final decision." Put qualitative / discussion.

Example of bare minimum for Qualitative / discussion: "resolve[d] disagreements by discussing a consensus annotation."

B.2.13. 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. This can also be in stating that in reconciliation, only items were kept if all annotators agreed (ex:(link removed)). Borderline in the other direction: (link removed) -- they discussed, but didn't report pre-discussion agreement.

- Yes
- No
- Unsure
- N/A -- if there was no multiple overlap

B.2.14. Total number of total human annotators who annotated items

How many total individuals were involved in evaluating/labeling items? Think of this like "number of annotators on the team." If it says there were initially 2 annotators, but a third joined halfway through, just put 3. This needs to be a single machine-readable number.

If you cannot answer the question because you don't know how many annotators (or it just says how many annotators per item), put a - and answer in the next question.

If they just say "we annotated", don't count the number of authors. Put no info.

If they didn't use human annotation, put N/A.

For surveys, total would be the number of respondents surveyed.

If all the info you have is an implicit "we labeled" (or other info that it was the authors), this is not sufficient information to answer the question.

B.2.15. Median number of human annotators per item

For papers that had multiple annotator overlap, when multiple annotators were involved in annotating the same item, how many annotated each item? Note that some papers will involve a smaller portion of items annotated by multiple annotators, and a larger portion annotated by just one. Exclude the cases where only one annotator annotated and determine the median number of annotators involved.

If there is no info on this (or it just says total number), put a -.

If they just say "we annotated", don't count the number of authors. Put no info.

Example: 75% of items labeled by 1 person, 20% labeled by 3 people, 5% labeled by 10 people. Put 3. This needs to be a single machine-readable number.

For surveys, median would be 1.

Example with bare minimum (physician's assessment): (link removed)

B.2.16. Link to dataset available:

Is there a link in the paper to the human-labeled dataset they used?

- Yes
- Yes, but link is broken
- No
- Unsure
- N/A if not using a human-labeled dataset

Only follow the link in the paper to see if it is broken, you don't need to verify it is actually a dataset.