A Rosbag Tool to Improve Dataset Reliability

Francesco Vigni
University of Naples Federico II
Naples, Italy
francesco.vigni@unina.it

Antonio Andriella
PAL Robotics
Barcelona, Spain
antonio.andriella@pal-robotics.com

Silvia Rossi
University of Naples Federico II
Naples, Italy
silvia.rossi@unina.it

ABSTRACT

Datasets are cornerstones of research in Human-Robot Interaction (HRI) and allow researchers to structure observations for other peers to work on. These often store temporal sensitive information about the behaviour of humans and robots involved in the study, and take advantage of the state of the art in robot logging, e.g., *rosbags*. Depending on the research goal, an approach commonly adopted is to publish datasets alongside annotated semantic information about the interaction. However, validating and assessing the quality of the datasets has not been the main concern of the community.

This work highlights the risk of publishing datasets without ensuring the synchronicity between objective and subjective measures and proposes a simple yet effective tool to mitigate it. The tool is evaluated on the *rosbags* of a popular dataset. Results show that 31.48% of its content contains indeterministic delays, causing the original synchronicity with the respective annotations to be lost.

CCS CONCEPTS

• **Applied computing → Digital libraries and archives**; • **General and reference → Reliability**.

KEYWORDS

Datasets validation, Failure taxonomy, Datasets reliability, Human-Robot Interaction.

ACM Reference Format:
Francesco Vigni, Antonio Andriella, and Silvia Rossi. 2024. A Rosbag Tool to Improve Dataset Reliability. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640556

1 INTRODUCTION

Building datasets in Human-Robot Interaction (HRI) is an impactful way of sharing research and progress as a community. There is, however, a lack of strict guidelines in this multidisciplinary field when producing and publishing datasets. Great insights can come from building datasets as collections of:

- **Objective measures:** These measures remain unaffected by personal opinions and encapsulate information systematically recorded during interactions, such as robot logs or sensor logs.
- **Subjective measures:** These can be swayed by the personal opinions of the rater and are frequently employed to annotate interactions with nuanced and complex information.

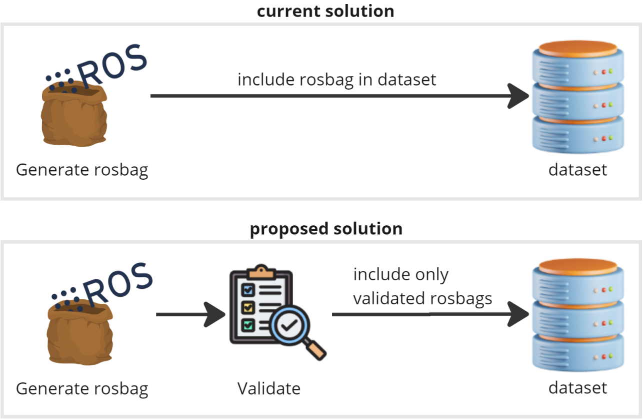


Figure 1: Proposed pipeline for validating *rosbags* before datasets inclusion.

When planning for the constructs of these measures, an approach that has demonstrated its advantages is to partially overlap subjective and objective measures [4, 12, 13]. The goal is to enable peers to exploit the produced dataset to address their own research questions while ensuring the highest possible quality. Any limitations in the dataset could potentially impact subsequent research endeavours. Additionally, sharing the data enhances reproducibility, an often overlooked but crucial aspect of HRI. In connection with this, consideration naturally turns to the methods used for assessing the quality of a dataset.

When considering subjective measures like annotations, a traditional approach is to calculate the inter-rater reliability of the annotators or coders along the data i.e. using Cohen's kappa coefficient or Correlation measures [11]. Regarding objective measures, little effort is invested in assessing the quality of the data, given that its inherent quality is intricately tied to the robot employed for its collection. Moreover, assessing the quality of an objective measure produced by a robot is a task that requires the researcher to manipulate robot logs in the form of binary files automatically generated and compressed by the robot. Due to the complexity of this task, the most common approach for including robot logs in a dataset is to include all the files automatically generated by the system (see *current solution* in Figure 1).

This approach relies on the software's capability to detect errors, notify the researcher and consequently stop the recording session without creating the log file. For instance, if the hard drive on which a robot log is about to be saved is broken or corrupted, the system halts the recording process.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640556

Despite this rationale, when collecting information from a real-world system like a robot, the software and its architecture can cause unforeseeable effects that can impact the data collection. This concern is particularly relevant in systems lacking real-time scheduling of processes, where internal processes can be paused and resumed without the researcher's control.

Furthermore, it is essential in HRI settings to ensure the alignment of subjective measures with objective ones. For example, an annotator might decide that a human and a robot are engaged only if they are both looking at each other. If this information is inaccurately included in the annotations (e.g., user and robot respectively gaze at that specific time), the reliability of the obtained dataset is questionable.

In this work, we focus on the risk of including an objective measure in a dataset in which the internal processes are not controlled. Additionally, we propose a simple tool to validate the quality of the objective measures collected using *rosbag* - a toolkit widely used by the community. We aim to safeguard datasets from potential pitfalls (see *proposed solution* in Figure 1) in light of improving their reliability.

2 BACKGROUND

Standardising how results are published may boost the progress of the field [6], as more researchers worldwide seek to replicate studies and benchmark their solutions on existing datasets ignoring potential quality issues.

In [15], Wienke *et al.* proposed a framework for the acquisition of multimodal HRI datasets. Their framework accounted for objective as well as subjective measures, however, the adoption of this approach is limited and requires integration with the event-based middleware named Robotics Service Bus (RSB) [16].

In [9], Lazzeri *et al.* developed a platform named HIPOP (Human Interaction Pervasive Observation Platform), designed for multimodal acquisition. HIPOP is a flexible system comprising diverse hardware and software components, enabling the configuration of personalized data collection setups for studies in HRI. By employing modules for capturing physiological signals, eye movements, video, and audio, the platform facilitates comprehensive analysis of both affective and behavioural aspects. Additionally, it allows for the integration of new hardware devices into the existing setup.

A step towards dataset reliability is presented in the Vernissage dataset [8] in which authors implement a post-processing mechanism to validate the synchronicity of all recorded data recorded with an RSB system.

Despite these works attempting to standardise how datasets are built in HRI while overcoming platform-tailored acquisition strategies, their adoption is still limited. This result is the product of choosing the rarely used RSB middleware while focusing on high-level data types such as physiological signals. To counter this, we investigate how the widely adopted middleware Robot Operating System (ROS)[1] can be used to build datasets, thanks to its own logging mechanisms, with a focus on low-level data types that can be recorded during an HRI.

2.1 The Role of ROS

ROS is the open-source standard *de facto* for building robot applications and is widely adopted in both academic and industrial settings. As of December 2023, ROS is used in at least 194 robots worldwide [17] and by 634 active companies [14].

It is shipped in two main versions: ROS1 and ROS2. Without delving into the technicalities of these nor their communication protocols, for this audience, it is important to highlight that the approach proposed in this work is particularly relevant for robots that are shipped with ROS1.

Regarding the newest version of the middleware, ROS2, the development is orbiting around how real-time constraints can be achieved within known extents. Therefore, the issues discussed in this manuscript are less relevant.

ROS comes equipped with a logging tool designed to generate files with *.bag*[2] extensions. These files, commonly referred to as *rosbags*, are acquired by selecting pertinent ROS topics within the system and are made of serialised message data published by these topics. These *rosbags* can also be played back in ROS, allowing researchers to include these files as objective measures of datasets. Despite this, a typical HRI dataset also contains subjective measures that are semantically or temporally linked to the objective ones, however, no prior work has focused on ensuring the reliability of these in light of the use cases offered by HRI.

3 TOWARDS RELIABLE DATASETS

Here, we first classify a taxonomy of failures when building datasets and propose a tool that can improve the reliability of datasets based on the time continuity constraint of the objective measures *rosbags*.

3.1 Failure taxonomy

Inspired by the failure taxonomy presented in [7], here, we classify failures in datasets design as *extrinsic* and *intrinsic* (see Figure 2).

Intrinsic failures are the ones encoded by ROS and commonly referred to as exceptions. These failures result in problems while creating a *rosbag* or while performing a playback, where the file is not properly created or cannot be properly read, respectively.

In contrast, *extrinsic* failures are the ones that do not result in problems when creating the file or when performing a playback, however, the semantics of the information associated with the *rosbag*, i.e. annotation, is compromised. For instance, when annotating with social labels messages from a *rosbag* in which an unforeseeable issue has occurred, i.e. network is overloaded, the time-dependent annotations will be stored with an uncontrolled time shift. The overall rationale is that when considering a reliable dataset the *semantic link* must be preserved. As a result, to avoid breaking it, hence mitigating *extrinsic* failures, the validation tool proposed here aims to classify *rosbags* into valid and invalid, employing a constraint on the temporal continuity of specific messages in the *rosbag*. This section does not present a complete taxonomy on the topic but drafts a simple one in light of the focus of this manuscript.

[1] https://www.ros.org/

[2] http://wiki.ros.org/Bags

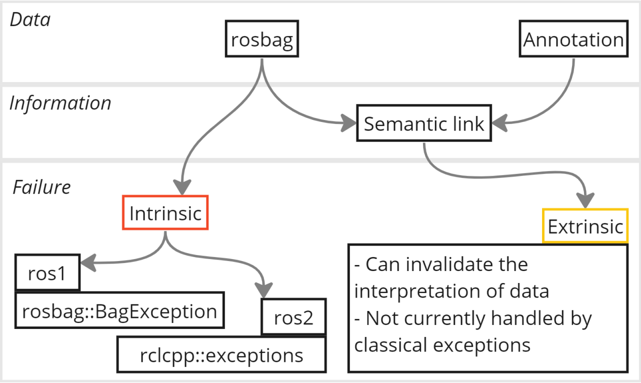


Figure 2: Failures' classification according to objective and subjective measures for dataset inclusion.

3.2 A Tool for Improving Reliability

The proposed tool analyses the ROS topics that are semantically important for the annotation phase and labels *rosbags* as valid only those for which the semantic link with the respective annotation is preserved. For example, if the annotations are obtained by the frame sequence of a camera, this tool expects a constant delay between each frame in order to label the related *rosbag* as valid.

On the other hand, a *rosbag* file is labelled as invalid when camera streams exhibit indeterministic delays (random freezing of camera frames) as this results in a misalignment with respect to its annotation. In this case, the intended social meaning stored in the subjective measures is lost, i.e., the semantic link is broken.

A reasonable metric for classifying a *rosbag* as either valid or invalid is the Standard Deviation (*std*) of consecutive ROS messages, such as camera frames. This approach allows us to evaluate the dispersion of data around their mean, yielding $std = 0$ for an ideal system. For all other cases, $std > 0$. The tool classifies as valid *rosbags* those for which the *std* is a reasonably small value, while the invalid ones are those for which *std* exceeded a threshold. Notice that we do not claim that the proposed metric *std* is optimal for the task, but explore it as a first attempt to build a validation tool. A software written in Python3.8 is publicly available[3] and uses GNU GPLv3 license with the following interface:

```
is_rosbag_valid(rosbag, topics, measure, thres) -> bool
```

4 A RUNNING EXAMPLE

Published in 2017, the UE-HRI dataset [2] provides the community with roughly 400GB of data in which the robot Pepper[4] was autonomously programmed to conduct social interactions, and collect data while deployed in a public hall. The bottom and front camera streams are annotated according to the following labels that encompass the social scene:

- Early sign of future engagement BreakDown (EBD) i.e. first noticeable clue that an engagement breakdown will occur in the remainder of the interaction.
- Engagement BreakDown (BD) i.e. leaving before the end of the interaction.
- Sign of Engagement Decrease (SED) observed during the interaction (None of the 3 next labels).
- Temporary disengagement (TD) i.e. leaving for some time and coming back to the interaction.

The dataset is published with *rosbags* (ROS1) for the objective measures and annotated ELAN files[5] for each *rosbag* regarding the subjective measures.

Each ELAN file stores time windows that are associated with the socially relevant labels mentioned above. The semantic link of this dataset is guaranteed if the camera streams (sources for the annotators) do not exhibit any indeterministic delays.

Hence, it is an ideal use case to evaluate the proposed tool. In an ideal condition, where the publishing rate is perfectly constant, frames would be periodically published at a known and fixed rate. Despite this, given the ROS1 limitations previously introduced, we might expect these streams not to publish frames at a constant rate. Figure 3 shows consecutive frames with respect to the timestamp of two different *rosbags* ("user104_2017-06-20" and "user106_2017-03-08") and includes the streams of the ROS topics used for the annotation phase:

- `/camera/front/image_raw`
- `/camera/bottom/image_raw`

named as *front* and *bottom* in the legend of the figures. These figures also highlight how the streams of the two analysed ROS topics are synchronised, producing overlapping plots. Importantly, notice the *steps* in Figure 3b that can explain at which timestamps the camera streams freeze. In this case, using such a file paired with the respective annotation results in breaking the semantic link between the objective (*rosbag*) and the subjective measure (ELAN file).

With the proposed solution it is possible to label as valid *rosbags*, those for which the *steps* are reasonably small, e.g., Figure 3a, and as invalid those that exhibit big *steps*. The urgency of this tool is manifested by the lack of strict guidelines for validating objective measures when building a dataset and by the result of its first evaluation on a popular dataset reported in the following section.

5 DISCUSSION

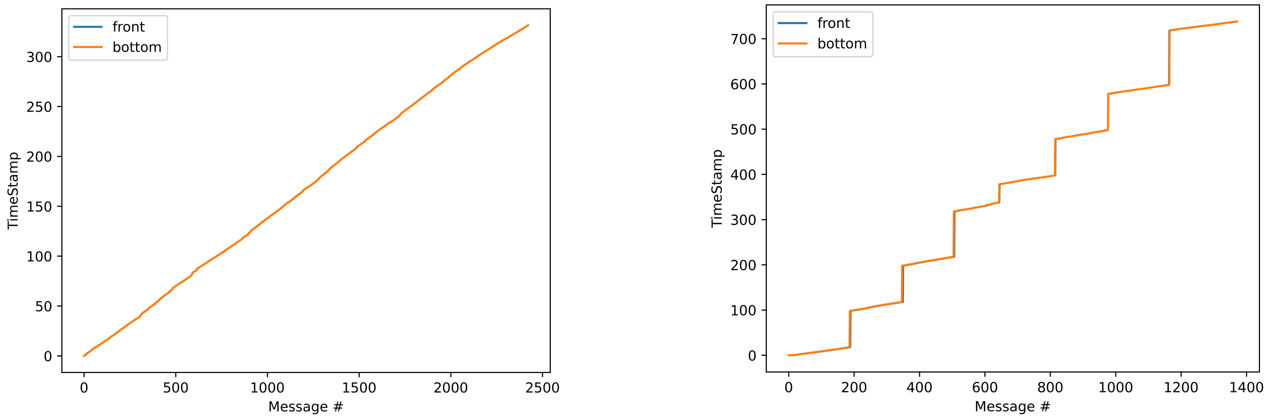
We tested the proposed tool on the *rosbags* of the UE-HRI dataset [2], which is the most widely-used dataset for machine learning in the HRI community [3, 5, 10]. We empirically set the threshold to 0.5 and studied the frames from the streams of the cameras *front* and *bottom* of the robot. Figure 4 summarises the standard deviation of each *rosbag*. Notice that for most of the *rosbags* the tool returned the metric (*std*) very close to zero (valid *rosbags*). This means that most of the content of the dataset maintains its semantic link, in other words, the temporal association with their respective annotation is preserved. On the contrary, 17 out of the 54 *rosbags* (31.48%) register a standard deviation higher than the set threshold. These are considered invalid *rosbags*, and shall not be used in pair with the respective annotation. It is also interesting to notice that only for a few *rosbags* the *front* and *bottom* camera streams show different standard deviations. The tool explores a safety first policy,

[3] https://github.com/Prisca-Lab/reliable-dataset

[4] https://www.aldebaran.com/en/pepper

[5] https://archive.mpi.nl/tla/elan

Figure 3: Snapshots of consecutive messages vs timestamps of two *rosbags* published in [2].



(a) Snapshot of rosbag "user104_2017-06-20".

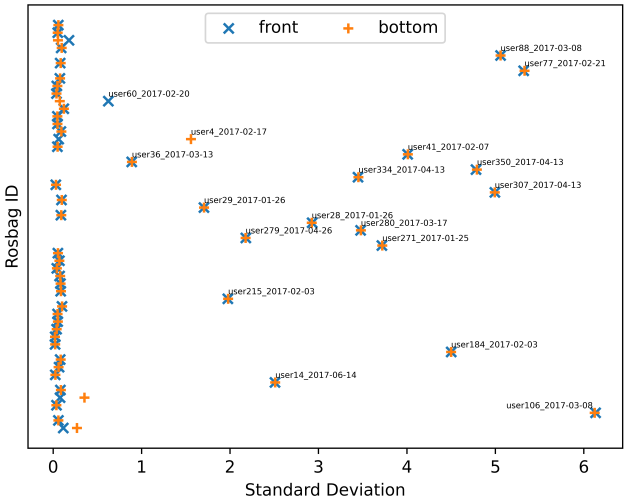
(b) Snapshot of rosbag "user106_2017-03-08".

meaning even if only one of the objective measures, i.e. camera streams of a *rosbag*, violates the set threshold, the whole sample (*rosbag*) is marked as invalid.

Together with filtering valid from invalid *rosbags*, the tool also allows us to understand how homogeneous a dataset is. For instance, if a peer is to use a dataset and manually inspect a few *rosbags*, the risk is that the randomly selected samples do not show any issues regarding frame freezing, leading to the use of the dataset assuming its consistency. However, if the task at hand is to train a machine-learning model with such a dataset, the common assumption is to have a homogeneous distribution of errors along the dataset. Unfortunately, as shown in this section, this is a weak assumption.

We also investigated if other datasets can benefit from this tool and concluded that in [1] and [13] the tool cannot be used since the authors deliver the dataset in raw data. The advantages introduced by this strategy are outmatched by the lack of standardization for manipulating raw data. In other words, peers who use raw data are more flexible in deciding how to process it, at the price of adopting heterogeneous strategies across the community.

Figure 4: Validation tool report of camera streams from [2].



6 CONCLUSIONS

This contribution highlights the importance of preserving the semantic link in the dataset between its objective and subjective measures. After drafting a taxonomy for failures when building datasets, this manuscript presents a tool that can mitigate the risk of *extrinsic* failures in terms of a time continuity constraint of its objective measures, i.e. *rosbags*.

Despite the preventative approach this tool aims at, here we show its usage as a validation tool on a publicly available dataset. A first version of the tool is implemented and evaluated on the popular dataset UE-HRI [2], and the results highlight that 31.48% of its *rosbags* are labelled as invalid. As a consequence, works that build upon this dataset have been using a partially valid source regarding the time synchronization between the *rosbags* and their annotations.

Future works will centre on improving the tool with real-time capabilities during *rosbag* recording for dataset creation. This advancement aims to empower researchers by providing immediate feedback on the validity of a *rosbag*, facilitating early error detection and the implementation of effective contingency strategies. Additionally, similarly structured datasets will be evaluated alongside other metrics than the presented Standard Deviation. The aim is to establish this tool as the standard method for validating robot logs produced by the large majority of existing robots, i.e., *rosbags*, to enhance the reliability of datasets in HRI.

ACKNOWLEDGMENTS

This work has been partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778 (F. Vigni), by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801342 (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ) (A. Andriella), and by the Italian Ministry for Universities and Research (MUR) with the PNRR Project FAIR (Future Artificial Intelligence Research) PE0000013 (S. Rossi).

REFERENCES

[1] Dustin Aganian, Benedict Stephan, Markus Eisenbach, Corinna Stretz, and Horst-Michael Gross. 2023. ATTACH Dataset: Annotated Two-Handed Assembly Actions for Human Action Understanding. *arXiv preprint arXiv:2304.08210* (2023).

[2] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[3] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics* 11 (2019), 815–828.

[4] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing* 10, 4 (2017), 484–497.

[5] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. 2020. Are you still with me? Continuous engagement assessment from a robot's point of view. *Frontiers in Robotics and AI* 7 (2020), 116.

[6] Hatice Gunes, Frank Broz, Chris S. Crawford, Astrid Rosenthal-von der Pütten, Megan Strait, and Laurel Riek. 2022. Reproducibility in Human-Robot Interaction: Furthering the Science of HRI. *Current Robotics Reports* 3, 4 (01 Dec 2022), 281–292. https://doi.org/10.1007/s43154-022-00094-5

[7] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[8] Dinesh Babu Jayagopi, Samira Sheikhi, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastian Wrede, Vasil Khalidov, Laurent Nguyen, Britta Wrede, and Daniel Gatica-Perez. 2012. *The vernissage corpus: A multimodal human-robot-interaction dataset.* Technical Report.

[9] Nicole Lazzeri, Daniele Mazzei, and Danilo De Rossi. 2014. Development and testing of a multimodal acquisition platform for human-robot interaction affective studies. *Journal of Human-Robot Interaction* 3, 2 (2014), 1–24.

[10] Tianlin Liu and Arvid Kappas. 2018. Predicting engagement breakdown in HRI using thin-slices of facial expressions. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

[11] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[12] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. 2013. Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3414–3421.

[13] David A Salter, Amir Tamrakar, Behjat Siddiquie, Mohamed R Amer, Ajay Divakaran, Brian Lande, and Darius Mehri. 2015. The tower game dataset: A multimodal dataset for analyzing social interaction predicates. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 656–662.

[14] Companies using ROS. 2023. https://github.com/vmayoral/ros-robotics-companies Last accessed: 2023-12-07.

[15] Johannes Wienke, David Klotz, and Sebastian Wrede. 2012. A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective. In *LREC 2012 Workshop on Multimodal Corpora for Machine Learning*. Citeseer.

[16] Johannes Wienke and Sebastian Wrede. 2011. A middleware for collaborative research in experimental robotics. In *2011 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 1183–1190.

[17] Robots with ROS. 2023. https://robots.ros.org/ Last accessed: 2023-12-07.