Dataset Documentation and Author Statement

This document concerns the Labor Inspection Checklist dataset (LICD) and paper submitted to Neurips2022 [1, 2]. The document is based on the «Datasheets for Datasets» and follows the recommended setup, according to the submission requirements for the conference. Many of the questions are indirectly answered in the main submission.

The supplementary materials include two Jupyter Notebook scripts and a copy of the Labour Inspection Checklists dataset, used for the experiment. The scripts are used to run the two demonstration experiments described in the main paper. Running the scripts requires Jupyter Notebook with Python 3.7 or higher. The dataset also needs to be located within the same folder, in order to run the script.

1 Motivation

For what purpose was the dataset created?

The data is collected as a part of NLIA's daily operations, where each inspection is digitally recorded and automatically added by their case management systems upon the completion of an inspection.

This dataset is created specifically for research on the use of machine learning for labor inspections, as stated in the main paper. The dataset could also possibly be used for benchmarking machine learning (ML) methods that addresses imbalanced or long-tailed target variables.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is collected by NLIA. This dataset is created by the main author of the paper (Eirik Flogard) who is employed as Senior Adviser at NLIA.

Who funded the creation of the dataset?

Not relevant. There are no conflicts of interests.

2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Each instance in the dataset is an inspection conducted by NLIA between 01/01/2012 and 01/06/2019. This is explained in further details in the paper.

How many instances are there in total (of each type, if appropriate)?

There are 63634 instances in the dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset contains every inspection conducted by NLIA within the time frame above. However, we excluded any records with organizations of less than 6 employees to be on the safe side in terms of privacy protection.

What data does each instance consist of?

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Each instance is an inspection that consists of target organization (many features), a checklist and a label indicating the outcome of the inspection. This is described in more details in the paper.

We would like to mention that each organization in the dataset has an industry code. There is a lookup table for the corresponding industry name for each code, which can be found at: `https://www.ssb.no/en/klass/klassifikasjoner/6`

Is there a label or target associated with each instance?

Yes, either the checklistsID or the non-compliance label can be regarded as a target. Another feature called "Checklist Content" lists the content of the checklist given by the ChecklistsID.

Is any information missing from individual instances?

The financial variables had missing information, which indicates that the organization did not report them. The missing values means that the value of the accounting post is 0.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

There are no explicit relations between individual instances in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)?

No.

Are there any errors, sources of noise, or redundancies in the dataset?

Not that we are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)?

No. The main author has gone through a review process with NLIA and taken precautions to not include any confidential or harmful information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Does the dataset relate to people?

No. As a rule of thumb, organizations are not subjected to privacy regulations in Norway because they are legal entities. To be on the safe side regarding privacy, any inspected organization in the dataset has at least 6 employees.

3 Collection Process

How was the data associated with each instance acquired?

The data is collected as a part of NLIA's daily operations, where each inspection is digitally recorded and automatically added by their case management systems upon the completion of an inspection.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

We used MSSQL17 (via Management Studio) to build the dataset from NLIA's databases.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Not relevant. The dataset is not a sample.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data is collected by the agency's inspectors and the databases is maintained by the agency. Everyone involved in the actual data collection and creation is employed at NLIA.

Over what timeframe was the data collected?

The data is formally inspections that have been completed between 01/01/2012 and 01/06/2019. The records of the inspected organizations are "snapshots" taken from the point in time when the inspection was conducted.

Were any ethical review processes conducted (e.g., by an institutional review board)?

We made sure that the dataset only contains information that are not harmful or violates any privacy protection laws. The content of the dataset has been reviewed by the management of NLIA and the main author.

Does the dataset relate to people?

No. The dataset is compliant to GDPR and Norwegian privacy protection laws.

4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The target label (Non-compliance) is constructed by assigning "1" to every inspection where at least 1 violation are found. "0" is assigned to every inspection where no violations are found. We decided to not include the exact number of violations found in the inspections as an extra precaution to prevent harm, in case single organizations in the dataset can be identified.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data is stored and maintained by NLIA's Datawarehouse developers, since the data is used for the agency's daily operations. In addition, a copy of the dataset at the time of creation is stored at NLIA's databases. These records are stored indefinitely by the agency.

Is the software used to preprocess/clean/label the instances available?

Yes, however access to NLIA's databases and the original 'raw' records is restricted.

5 Uses

Has the dataset been used for any tasks already?

We have used the dataset for a feature selection experiment in our previous work [3], without publishing the dataset.

We have also described two possible use cases for the dataset in the main paper [2] (see the demonstration experiments). The first case is selecting a relevant labour inspection checklist for a given organization. The second case is to predict whether an organization is non-compliant to working environment regulations, where selected checklists can be used as independent features.

Is there a repository that links to any or all papers or systems that use the dataset?

No.

What (other) tasks could the dataset be used for?

The dataset could be relevant for SDG practitioners who want to build ML models for promoting decent work. We have also provided some directions for potential uses in the "Conclusion and Future work" section of the main paper [2].

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No.

Are there tasks for which the dataset should not be used?

No.

6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

The dataset will be openly accessible to the public.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset will be published in Dataverse under the following DOI:
https://doi.org/10.18710/7U6TZP.

When will the dataset be distributed?

The dataset will be published upon acceptance of the paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset will be made publicly available. This dataset is licensed under a CC0 License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

7 Maintenance

Who is supporting/hosting/maintaining the dataset?

The host of the dataset is the main author, who works as senior adviser at the Norwegian Labour Inspection Authority.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The manager can be reached at Eirik.flogard@arbeidstilsynet.no.

Is there an erratum?

No, any corrections will be made to the dataset if needed.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

The dataset will be updated when necessary (for instance correcting errors) or by request from users (new instances). New versions of the dataset will be labeled appropriately.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

Not relevant.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, if it is relevant to do so.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset is published openly under a CC0 license. We will also gladly accept contributions or updates to the dataset. The procedure for this would be to contact the main author by email and present the contribution in question.

8 Author Statement

The dataset is licensed under CC0 License. The main author confirms that the dataset adhere to privacy laws and ethical standards and will bear all responsibility in case of violation of such laws, rights and standards.

Sign.

……………………………………………………………

Eirik Lund Flogard

References

[1] Eirik Lund Flogard. Labour Inspection Checklist Dataset. `https://doi.org/10.18710/7U6TZP`, 2022.

[2] Eirik Lund Flogard and Ole Jakob Mengshoel. A dataset for efforts towards achieving the sustainable development goal of safe working environments. In *Advancements in Neural Information Processing Systems, Data set and Benchmark Track*, 2022.

[3] Ole Jakob Mengshoel, Eirik Flogard, Jon Riege, and Tong Yu. Stochastic local search heuristics for efficient feature selection: An experimental study. In *Norsk IKT-konferanse for forskning og utdanning*, number 1, pages 58–71, 2021.