Engineering a Machine Learning Pipeline for Automating Metadata Extraction from Longitudinal Survey Questionnaires

Suparna De
Department of Computer Science, University of Surrey

30th November 2021

on behalf of:
Harry Moss, Sanaz Jabbari: Centre for Advanced Research Computing, UCL
Haeron Pereira: Department of Computer Science, University of Surrey
Jon Johnson, Jenny Li: CLOSER, UCL Institute of Education

13th Annual European DDI User Conference
30th November – 1st December 2021, Online (Paris)

Presentation Outline

- Background
- Problem Statement
- Proposed Approach
- CLOSER ML pipeline
- Model Experiments and Results
- Conclusions

Background

- DDI-Lifecycle
  - robust metadata model for questionnaire content and flow capture
  - support for versioning and provenancing objects
- Archives
  - information in PDFs associated with surveys
- CLOSER Discovery
  - a range of social sciences and biomedical domains' longitudinal studies
  - provision of questionnaire metadata in DDI-Lifecycle → manually/semi-manually

Problem Statement

- Increased volume of data → more questionnaires added to CLOSER Discovery
- Scaling to provide a high-quality question bank
- Survey questions for reuse by studies and data collection agencies → reproducibility of studies and analyses

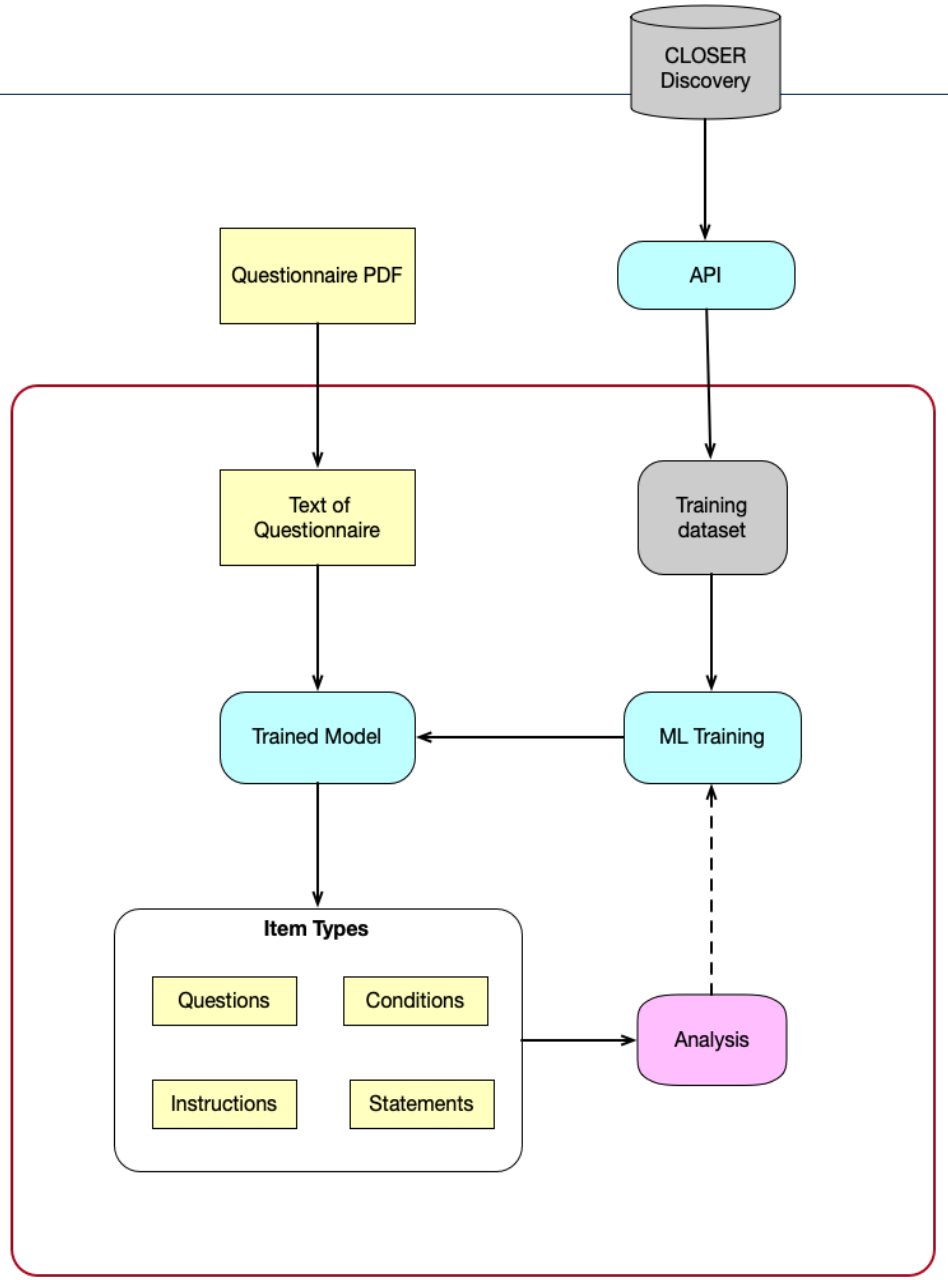
Ease, efficiency and robustness of metadata extraction of question items

Automated methods for questionnaire item metadata extraction

Proposed Approach

- Supervised learning algorithms
  - candidate approach for automating the extraction of valid DDI items from the survey questionnaires in PDF format
  - training and validation dataset - processed and marked-up (in XML) questionnaires
  - text classification problem
- Continuous build and integrate approach: ML pipeline with combinations of:
  - input data,
  - feature engineering methods,
  - model parameters
  - resultant outputs
- Experiments' metadata
  - various combinations experimented with, and the corresponding outputs

  → reproducibility, comparative analysis and provenance of the pipelines

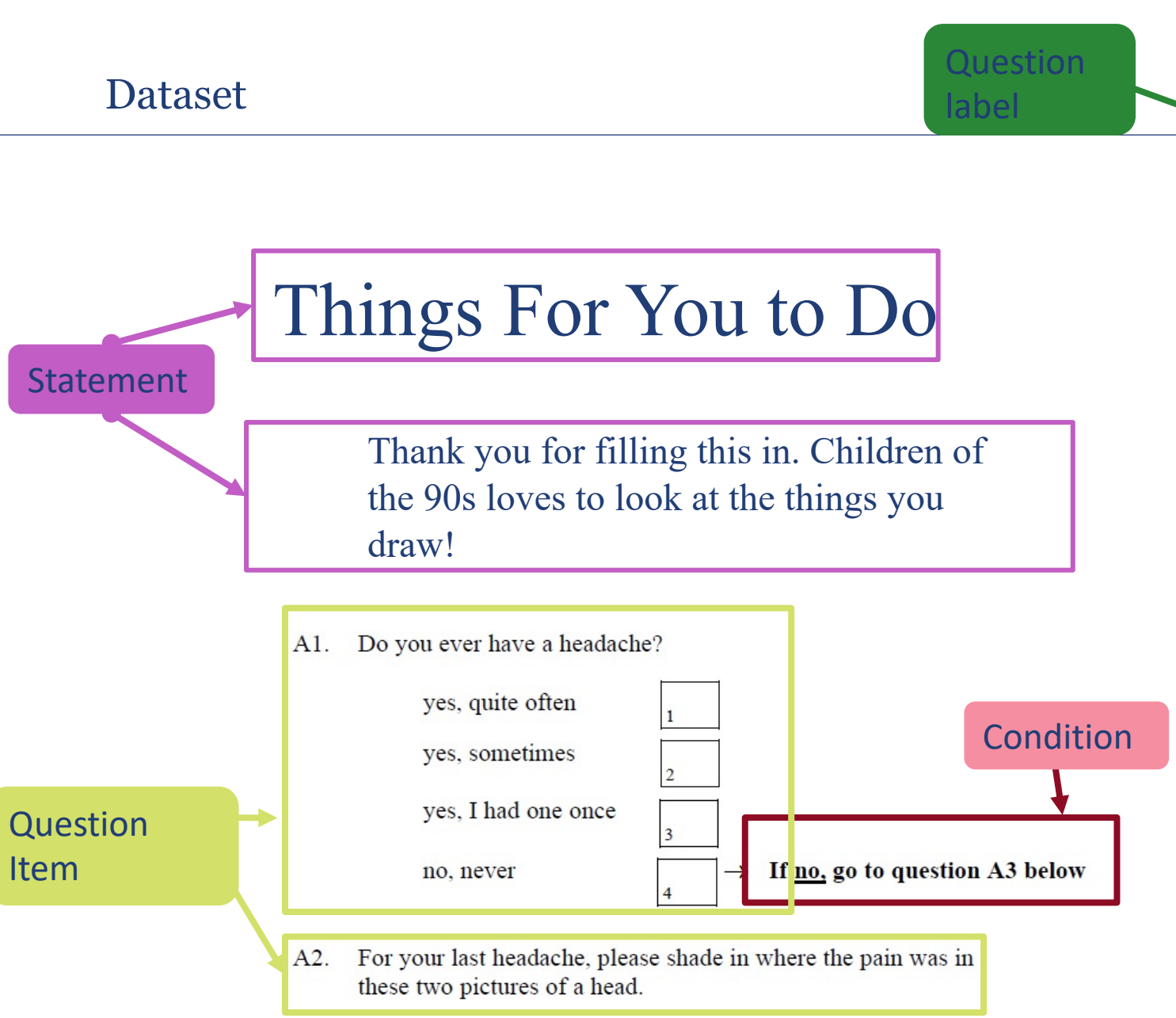


Proposed Approach (2)

- Abstraction of model parameters in pipelines
  - Data Version Control (DVC)

- Automating the process of attaching metadata related to each model experiment
  - ProvLake

Dataset

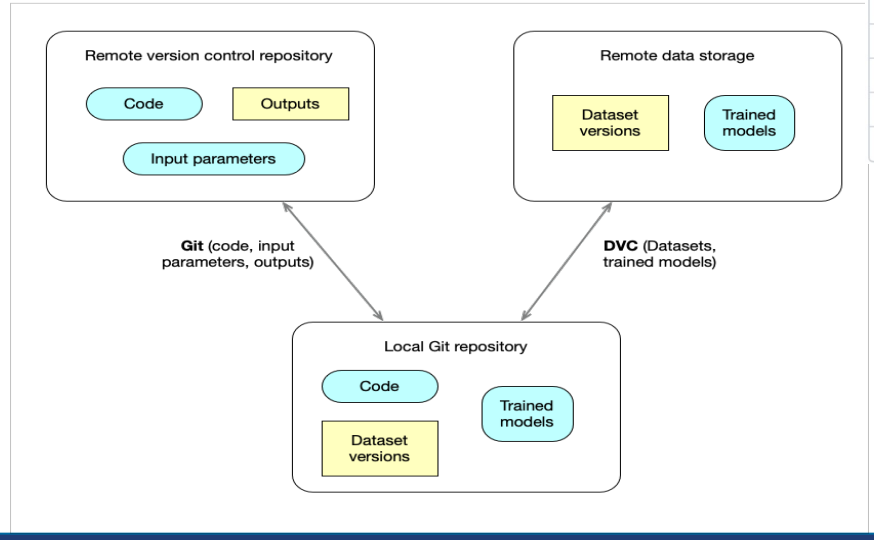


Item Type	Count
QuestionName	40,546
Question literal	40,545
Interviewer Instruction	3,051
Statement	7,551
Response Domain - Codelist	85,547
Response Domain - DateTime	486
Response Domain - Text	613
Conditional	8,414
Loop	352
Total	187,105

CLOSER ML pipeline

- DVC
  - dataset versioning and experiment tracking
  - version the state of the input data as it evolves
  - associate that with a specific git commit hash
  - transfer the versioned data over SSH to remote storage

- Remote data storage: UCL RDSS
  - petabyte-scale storage facility



multinomial... 9 branches 9 tags Go to file Add file Code

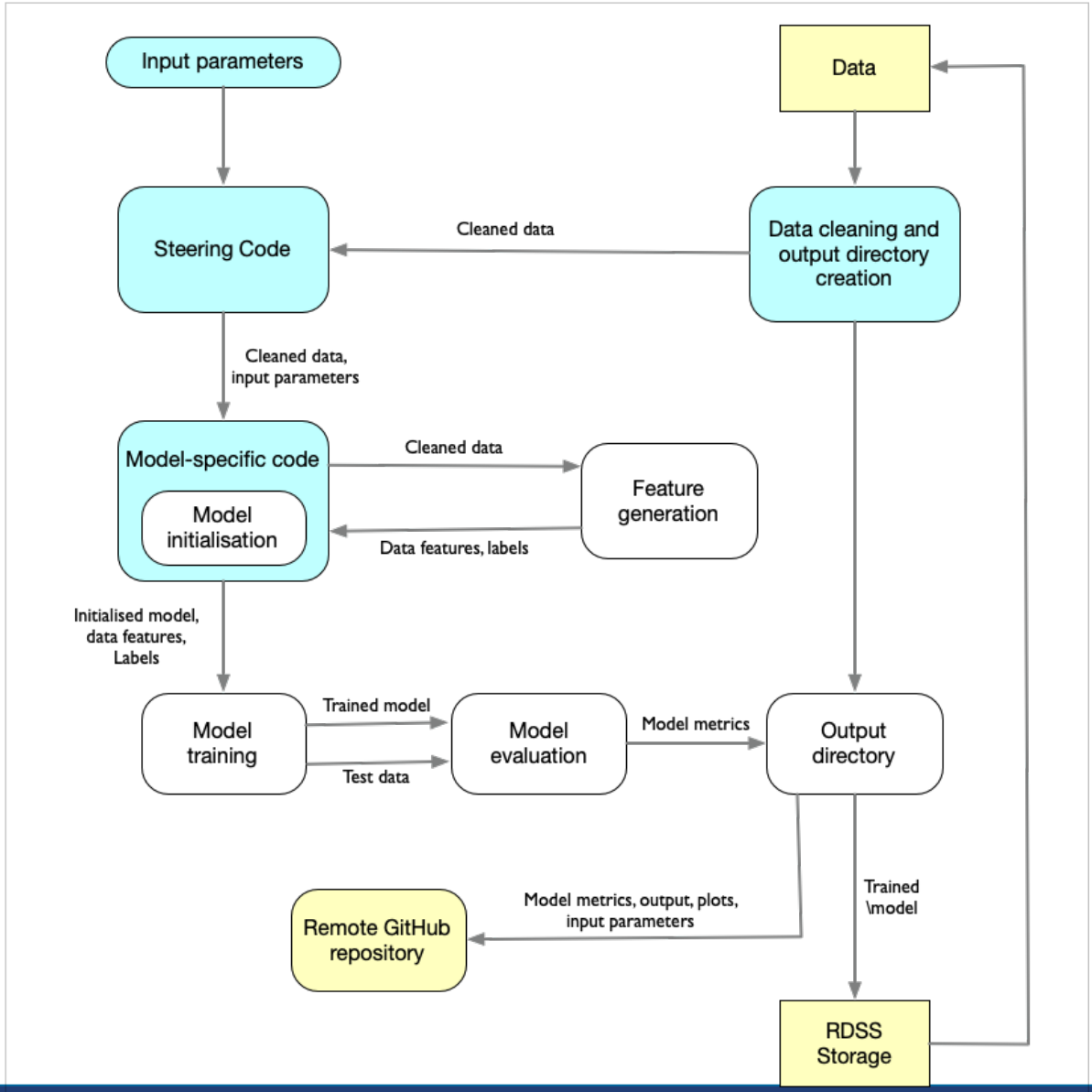
This branch is 49 commits ahead of main. Contribute

harryjmoss Merge branch 'main' into multinomi... 90a3e84 10 days ago 137 commits

.dvc	Configure RDSS remote	9 months ago
.github/workflows	Add basic MLFlow parameter tracking (#76)	13 days ago
data	Adding additional ESRC datasets	10 days ago
myriad	Type hinting and small refactor (#75)	last month
notebooks	Add naive-bayes large dataset notebook	8 months ago
output	Adding multinomial NB esrc re-run	4 months ago
requirements	Update model code in main (#77)	12 days ago
src	Inference fixes for BERT models	10 days ago
.dvcignore	Initialise DVC	9 months ago
.gitignore	BERT base uncased (#28)	6 months ago
Dockerfile	first commit	10 months ago
LICENSE	first commit	10 months ago
README.md	Update README.md	12 days ago
dvc.lock	Adding multinomial NB esrc re-run	4 months ago
dvc.yaml	Adding multinomial NB esrc re-run	4 months ago
environment_setup_job.sh	Initial commit	10 months ago
hyperparam_optimisation...	Add basic MLFlow parameter tracking (#76)	13 days ago
mypy.ini	Type hinting and small refactor (#75)	last month
params.yaml	Merge branch 'main' into multinomial_naive_bayes	13 days ago
run.py	Add basic MLFlow parameter tracking (#76)	13 days ago
run_rcnic_job.sh	Update model code in main (#77)	12 days ago
setup.cfg	Add testing for data prep pipeline stages (#15)	7 months ago
setup_spacy.sh	Adding all three base model types and updates to...	7 months ago

Model Experiments

- Model
  - Multinomial Naïve Bayes
- Experiment setup
  - Python 3
  - pyTorch, scikit-learn ML libraries
- Model metrics
  - Evaluation metrics
    - accuracy, precision, recall, f1-score
    - AUC - ROC curve
    - Confusion matrix
  - Hyperparameter tuning
    - Gridsearch



Model Results


```
[
    {
        "prov_obj": {
            "task": {
                "id": 1633622500.4462595,
                "wf_execution": 1633622400.2880254,
                "startTime": 1633622500.4462595,
                "generatedTime": 1633622500.4462595,
                "status": "RUNNING"
            },
            "dt": "GridSearchCV",
            "type": "Input",
            "values": {
                "priors": [
                    null
                ],
                "var_smoothing": [
                    1e-08,
                    ...,
                    0.00012
                ]
            }
        },
        "dataflow_name": "Naive Bayes Model",
        "act_type": "task"
    }
]
```

```
[
    {
        "prov_obj": {
            "task": {
                "id": ,
                "wf_execution": ,
                "startTime":
                "endTime":
                "generatedTime":
                "status": "FINISHED"
            },
            "dt": "GridSearchCV",
            "type": "Output",
            "values": {
                "f1score": [
                    ...
                ],
                "mean_test_accuracy": [
                    ...
                ],
                "mean_test_recall": [
                    ...
                ],
                "mean_test_precision": [
                    ...
                ]
            }
        },
        "dataflow_name": "Naive Bayes Model",
        "act_type": "task"
    }
]
```

Conclusions

- Automation of data and metadata extraction from longitudinal survey questionnaires through:
  - supervised ML pipeline approach
- Challenges in ML pipelines
  - data version control
  - managing changes to models and datasets
- Reproducible ML model training and execution method
  - generates logging metadata in a structured format
  - tracking of various combinations of
    - input data,
    - model features
    - hyperparameter tuning with
    - obtained output values
- DDI-Lifecycle schema
  - rich provenance structure allows the analysis of prediction of specific item types (e.q. question text)

Thank you

Thank you!

Suparna De
s.de@surrey.ac.uk

http://metadata-automation.org

License