From UI designs image to GUI skeleton: A neural machine translator to bootstrap mobile GUI implementation Paper

ID

Experiments

AP1
4
The first experiment is model hyperparameters fine-tuning. No info is given about it. Could be embedded in E2 or : Comments

| Aspect | Element | E1 | Comments |
|------------------------|---|--|--|
| Experiment type | | Optimization | Model hyperparameters fine-tuning. |
| | | | "Hidden" except beam-width that appears |
| I bornathanana | December to the second | No | with E2 |
| Hypotheses | Research hypotheses Statistical hypotheses | No No | |
| Variables selection | Model hyperparameters | #layers: 14 (?) | Hidden states in LSTM=256. Filter size=3, |
| variables selection | Wodernyperparameters | connections: (convolutional + pooling)x6 + | stride=1 and zero pading=2 for conv layers, |
| | | RNN encoder (LSTM) + RNN decoder | first conv layer=64 filters (subsequents x2), |
| | | #neurons/layer: Not mentioned | pooling units=2x2 and stride=2 for pooling |
| | | The state of the s | |
| | | activation functions: convolutional (ReLU), | layers |
| | | pooling (?), encoder (?), decoder (?) | "Charles " Cl "Daniels |
| | | params. Initialization: No | #CNN layers, #conv layers filers, #RNN hidder |
| | | | states fine-tuned using another randomly |
| | | 11. | selected 3% of dataset |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Partially | Unknown size of input |
| | | model type: Yes | |
| | | loss function: Yes | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split:Yes | 90%-3% |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 10 | |
| | Training data | Yes | 90% Android UI dataset |
| Operationalization | Factors and treatments | No | Seems #CNN layers, #conv layers filers, #RNN |
| | | | hidden states |
| | | | Beam-width is explicitly mentioned (15) |
| | Response variable, elaboration and metric | No | Unknown |
| Design | Design type | No | |
| Design. | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | | | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Mentions average BLEU score (?) |
| Instrumentation | Test set | 3% set | Same as training. Can be downloaded |
| | Measuring instruments | No | 0 |
| | Measurement procedure | No | |
| | Pre-processing | Yes | Not needed |
| | Dataset construction | Yes | Stoat (run on 64-bit Ubuntu 16.04 server with |
| | | | 32 Intel Xeon CPUs, 189GB RAM, controls 16 |
| | | | emulators in paaralleleach app run for 45 |
| | | | mins.), soot, dexpler |
| | Technological infrastructure | Yes | Torch framework written in Lua |
| | - | | Nvidia M40 GPU (24GB memory) |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Partially | Not exactly clear the samples |
| Population | Objects (chars. of the experimental datasets) | Partially | Not exactly clear the samples |
| | | Partially | |
| | Objects (chars. of the experimental datasets) Descriptive statistics | · | Not exactly clear the samples No results at all are provided from this experiment |
| Population Analysis | Descriptive statistics | No | No results at all are provided from this |
| Analysis | | · | No results at all are provided from this |
| | Descriptive statistics Inferential statistics | No No | No results at all are provided from this |

From UI designs image to GUI skeletc

Paper ID Experiments

AP1
4
The first experiment is model hyperpæimply missing. E3 (generalization using other test sets). Just assesses h Comments

| Aspect | Element | E2 | Comments |
|---------------------------------|---|-------------------------------------|---|
| Experiment type | | Evaluation | Does not really compare against anything |
| | | | |
| Hypotheses | Research hypotheses | No | |
| nypotrieses | Statistical hypotheses | No No | |
| Variables selection | Model hyperparameters | #layers: 14 | Hidden states in LSTM=256. Filter size=3, |
| | | connections: (convolutional + | stride=1 and zero pading=2 for conv layers, |
| | | pooling)x6 + RNN encoder (LSTM) + | |
| | | RNN decoder | pooling units=2x2 and stride=2 for pooling |
| | | #neurons/layer: Not mentioned | layers |
| | | activation functions: convolutional | |
| | | (ReLU), pooling (?), encoder (?), | Not sure if these are the definite values |
| | | decoder (?) | obtained fom the previous experiment. |
| | | params. Initialization: No | Except beam-width that 2 is chosen |
| | Model parameters | Same as previous | Except bearn-width that 2 is chosen |
| | Woder parameters | Same as previous | |
| | DL algorithm | Same as previous? | Since E1 and E2 are merged, we can deduce |
| | | | the values |
| | | | |
| | Training hyperparameters | Same as previous | 90%-7% |
| | Training Hyperparameters | Same as previous | 30,07,0 |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes (deduced) | One treatment only |
| | | , | , |
| | | | |
| | Response variable, elaboration and metric | Yes | Accuracy: exact match rate, BLEU |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | Yes (deduced) | Depth of component hierarchy (39), #GUI |
| | | | coponents (565 step 5), @containers |
| | | | (324 step 3). Only for beam-width=2 |
| | Randomization | No | |
| | Task duration | ??? | Trained for 4.7 days (inluding E1?) |
| | Number of experimental units | No | |
| Instrumentation | Test set | 7% set | Same as training. Can be downloaded |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Pre-processing | Yes | |
| | Dataset construction | | |
| | | | |
| | Technological infrastructure | Yes | Torch framework written in Lua |
| 5 10 | | B. C. II | Nvidia M40 GPU (24GB memory) |
| Population | Objects (chars. of the experimental datasets) | Partially | Not exactly clear the samples |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | | No | |
| Validity evaluation | Conclusion, internal, construct, external | | |
| Validity evaluation Artifact | Availability | Yes | |

Paper From UI designs image to GUI skeletc

ID AP1 Experiments

4
The first experiment is model hyperpaow good it is. Comments

| Poper Pope | Aspect | Element | E3 | Comments |
|--|---------------------|---|-------------------|---|
| Variables selection Model hyperparameters Same as previous | | | | |
| Statistical hypotheses Variables selection Model hyperparameters Model parameters DL algorithm Same as previous Paper reads "we train our model". We can assume the same single training is used for all experiments. Training hyperparameters Same as previous Training data Operationalization Factors and treatments Pesign Type Response variable, elaboration and metric Design Design Design type Response variables Held-constant variables No No Design Design We train our model". We can be deduced Could be app??? 20 Orangetely unseen apps Fartally Can be deduced Partally Yes Ves Ves Ves Ves Ves Ves Ves | | | | |
| Statistical hypotheses Variables selection Model hyperparameters Same as previous? Beam-widle-2 is explicitly mentioned Model parameters Same as previous DL algorithm Same as previous Paper reads "we train our model". We car assume the same single training is used for all experiments. Training hyperparameters Same as previous Training data Same as previous Operationalization Factors and treatments Yes (Seduced) One treatment only Perhaps it is app? Response variable, elaboration and metric Yes Accuracy: exact match rate, BLEU Design Design type No Deduced Could be app??? 20 Measured variables No Deduced Could be app??? 20 Measured variables (covariates) No | Hypotheses | Research hypotheses | No | |
| Model parameters DL algorithm Same as previous DL algorithm Same as previous Paper reads "we train our model". We car assume the same single training is used for all experiments. Training hyperparameters Same as previous Training data Training data Operationalization Factors and treatments Response variable, elaboration and metric Design Design type Blocking variables Held-constant variables Measured variables (covariates) No Randomization Task duration No No 122 Partially Active Accuracy: exact match rate, BLEU No 123 Design type Blocking variables No No Randomization Task duration No No 124 Partially Active Accuracy: exact match rate, BLEU No 125 Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Yes Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Ves No | ** | | | |
| DL algorithm Same as previous Paper reads "we train our model". We can assume the same single training is used for all experiments. Training hyperparameters Same as previous Training data Operationalization Pactors and treatments Personse variable, elaboration and metric Design Design type Blocking variables Held-constant variables Measured variables (covariates) No No Randomization Task duration No Number of experimental units Measurement procedure Pre-processing Dataset construction No No Number of experimental with partially Measurement procedure Pre-processing Dataset construction Population Objects (chars, of the experimental datasets) Partially Partially Partially Partially Partially Objects (chars, of the experimental datasets) Partially Partially Partially Pres Population Objects (chars, of the experimental datasets) Partially Pres Population Objects (chars, of the experimental datasets) Partially Population Objects (chars, of the experimental datasets) Population Objects (chars, of the experimental datasets) | Variables selection | Model hyperparameters | Same as previous? | Beam-widt=2 is explicitly mentioned |
| DL algorithm Same as previous Paper reads "we train our model". We can assume the same single training is used for all experiments. Training hyperparameters Same as previous Training data Operationalization Pactors and treatments Personse variable, elaboration and metric Design Design type Blocking variables Held-constant variables Measured variables (covariates) No No Randomization Task duration No Number of experimental units Measurement procedure Pre-processing Dataset construction No No Number of experimental with partially Measurement procedure Pre-processing Dataset construction Population Objects (chars, of the experimental datasets) Partially Partially Partially Partially Partially Objects (chars, of the experimental datasets) Partially Partially Partially Pres Population Objects (chars, of the experimental datasets) Partially Pres Population Objects (chars, of the experimental datasets) Partially Population Objects (chars, of the experimental datasets) Population Objects (chars, of the experimental datasets) | | | | |
| Training hyperparameters Training data Operationalization Factors and treatments Operationalization Response variable, elaboration and metric Obesign Design type Blocking variables Held-constant variables Held-constant variables Held-constant variables Held-constant variables Held-constant variables No Measured variables (covariates) No | | Model parameters | Same as previous | |
| Training data Operationalization Factors and treatments Pes (deduced) Operationalization Factors and treatments Yes (deduced) One treatment only Perhaps it is app? Response variable, elaboration and metric Pes Accuracy: exact match rate, BLEU No Design Design type Blocking variables Held-constant variables Measured variables (covariates) No Measured variables (covariates) No No Task duration Task duration No No No Test set Measuring instruments Measuring instruments Measuring instruments Measuring instruments Partially Measuring instruments Partially Can be deduced Partially C | | DL algorithm | Same as previous | Paper reads "we train our model". We can assume the same single training is used for all experiments. |
| Operationalization Factors and treatments Yes (deduced) One treatment only Perhaps it is app? Response variable, elaboration and metric Yes Accuracy: exact match rate, BLEU Design Design type Blocking variables Deduced Could be app??? 20 Held-constant variables No Measured variables (covariates) No No Task duration No No No Measured variables (covariates) No No Task duration No No No Measured variables (covariates) No Test set Measuring instruments Partially 20 completely unseen apps (Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Yes Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Yes Inferential statistics Yes Inferential statistics No | | Training hyperparameters | Same as previous | |
| Operationalization Factors and treatments Yes (deduced) One treatment only Perhaps it is app? Response variable, elaboration and metric Yes Accuracy: exact match rate, BLEU Design Design type Blocking variables Deduced Could be app??? 20 Held-constant variables No Measured variables (covariates) No No Task duration No No No Measured variables (covariates) No No Task duration No No No Measured variables (covariates) No Test set Measuring instruments Partially 20 completely unseen apps (Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Yes Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Yes Inferential statistics Yes Inferential statistics No | | Training data | Same as provinus | |
| Response variable, elaboration and metric Perhaps it is app? Accuracy: exact match rate, BLEU Design type Blocking variables Held-constant variables Measured variables (covariates) No Randomization Task duration No Number of experimental units No Number of experimental units No Test set Measuring instruments Measuring instruments Measuring instruments Partially Pre-processing Dataset construction Technological infrastructure Population Objects (chars. of the experimental datasets) Descriptive statistics Analysis Descriptive statistics No Validity evaluation Conclusion, internal, construct, external Ardifact Avallability Yes Accuracy: exact match rate, BLEU Could be app??? 20 Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Accuracy: exact match rate, BLEU Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Accuracy: exact match rate, BLEU Could be app??? 20 Accuracy: exact match rate, BLEU Accuracy: exact match r | Operationalization | | | One treatment only |
| Design Design type Blocking variables Deduced Could be app??? 20 Randomization Task duration No | -, | | 100 (0000000) | |
| Blocking variables Held-constant variables Held-constant variables Measured variables (covariates) Randomization Task duration No Number of experimental units No Neasuring instruments Measuring instruments Measurement procedure Partially Pre-processing Dataset construction Technological infrastructure Population Objects (chars. of the experimental datasets) Analysis Descriptive statistics Validity evaluation Inferential statistics No Validity evaluation Occupancy Validity evaluation Volution Validity evaluation Conclusion, internal, construct, external No Artifact Avaliability Validity Ves | | Response variable, elaboration and metric | Yes | Accuracy: exact match rate, BLEU |
| Held-constant variables Measured variables (covariates) Randomization Task duration No Number of experimental units No Instrumentation Neasuring instruments Measuring instruments Measurement procedure Pre-processing Dataset construction Technological infrastructure Population Objects (chars. of the experimental datasets) Analysis Descriptive statistics Validity evaluation Inferential statistics Validity evaluation No No No No No No No Validity evaluation No No Validity evaluation Conclusion, internal, construct, external No Artifact Avaliability No | Design | Design type | No | |
| Randomization No No No No No No No | | Blocking variables | Deduced | Could be app??? 20 |
| Randomization Task duration No No Number of experimental units No 1?? Instrumentation Test set Partially Measurement procedure Pre-processing Dataset construction Technological infrastructure Population Objects (chars. of the experimental datasets) Analysis Descriptive statistics Inferential statistics Ves Validity evaluation No Volidity evaluation No Volidity evaluation No Validity evaluation Conclusion, internal, construct, external Ardifact Avallability No Ves Inferential statistics No Ves | | Held-constant variables | No | |
| Task duration No 17? Instrumentation Test set Partially 20 completely unseen apps Measuring instruments Partially Can be deduced Partially Pre-processing Partially Pres Pre-processing Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | | Measured variables (covariates) | No | |
| Task duration No 1?? Instrumentation Test set Partially 20 completely unseen apps Measuring instruments Partially Can be deduced Partially Pre-processing Pre-processing Same as previous? Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | | Randomization | No | |
| Number of experimental units No 1?? | | | | |
| Instrumentation Test set Measuring instruments Partially Can be deduced Partially Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Partially Can be deduced Partially Can be deduced Partially Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Can be deduced Partially Partially Can be deduced Partially Partially Can be deduced Partially Can be deduced Partially Partially Can be deduced Part | | | | 1?? |
| Measuring instruments Measurement procedure Pre-processing Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Population Objects (chars. of the experimental datasets) Inferential statistics Ves Validity evaluation Conclusion, internal, construct, external Availability Ves Ocan be deduced Can be deduced Can be deduced Can be deduced Can be deduced Can be deduced Ves | Instrumentation | | | |
| Measurement procedure Pre-processing Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Population Population Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Yes Inferential statistics Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | | | | |
| Pre-processing Dataset construction Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Partially Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | | | | Can be deduced |
| Technological infrastructure Same as previous? Population Objects (chars. of the experimental datasets) Population Analysis Descriptive statistics Inferential statistics Validity evaluation Conclusion, internal, construct, external Artifact Availability Artifact Availability Same as previous? Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Yes No Validity evaluation Yes | | | | |
| Population Objects (chars. of the experimental datasets) Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external Artifact Availability Partially 20 randomly choosen apps that have at least 1 million installations. They are not in previous dataset Yes | | Dataset construction | | |
| least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external Artifact Availability Yes | | Technological infrastructure | Same as previous? | |
| least 1 million installations. They are not in previous dataset Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external Artifact Availability Yes | | | | |
| Analysis Descriptive statistics Yes Inferential statistics No Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | Population | Objects (chars. of the experimental datasets) | Partially | least 1 million installations. They are not in |
| Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | Analysis | Descriptive statistics | Yes | previous dataset |
| Validity evaluation Conclusion, internal, construct, external No Artifact Availability Yes | | Inferential statistics | No | |
| Artifact Availability Yes | Validity evaluation | | | |
| | | | | |
| | | Badge | | |

From UI designs image to GUI skeletc

Paper ID Experiments

AP1 4 The first experiment is model hyperpa Comments

| Aspect | Element | E4 | Comments |
|---------------------------------|---|-------------------|---|
| Experiment type | | Generalization | Used by developers. Start from |
| | | | scratch/usign DNN output |
| | | | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | start generating GUIs from scratch or using the output of the DNN |
| | Response variable, elaboration and metric | Yes | time, user satisfaction (5-likert), expert judgement of similarity (5-likert) |
| Design | Design type | No | seems 1 factor |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | Partially | |
| | Number of experimental units | Yes | 8 people (each implementing same 5 UI images |
| Instrumentation | Test set | Partially | 5 UI images in Android (does not say which one |
| | Measuring instruments | Partially | Can be deduced |
| | Measurement procedure | Partially | Can be deduced |
| | Pre-processing Dataset construction | Yes | |
| | Technological infrastructure | Same as previous? | |
| Population | Objects (chars. of the experimental datasets) | Partially | PhD students |
| Analysis | Descriptive statistics | Partially | On some variables only |
| | Inferential statistics | Yes | Non-parametrics used |
| | | | Pro Control Control |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | No Yes | |

Paper Deep code search

ID AP2 Experiments 1

| Commen | ıts |
|--------|-----|
| | |

| Aspect | Element | E1 | Comments |
|---------------------|---|---|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Input, L1: bi-directional LSTMx3 + | |
| | | MLP; L2: max pooling x 4; L3:MLP; Output?? | |
| | | #neurons/layer: Input: 100, LSTMs: 200, MLF | |
| | | L1:100, MLP L3: 400, maxpooling:? | |
| | | connections: Yes | |
| | | activation functions: tahn | |
| | | params. Initialization: No | |
| | Mandal managements | Marana Ma | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: ranking loss | |
| | | regularization: No | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: No | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: 128 | |
| | | #epochs: 500 | |
| | Training data | Yes | Data available in github |
| Operationalization | Factors and treatments | Model type (CodeHow, Lucene, proposal) | CodeHow and Lucene are not DNNs |
| o perationalization | Response variable, elaboration and metric | Yes | FRank, Success-rate@k, Precision@k, MRR. |
| | nesponse variable, claboration and metric | 103 | Described in detail |
| Design | Design type | No | Described in detail |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Data available in github |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | Keras, Theano, Nvidia K40 GPU | Missing info about SO, versions, etc. |
| Population | Objects (chars. of the experimental datasets) | Partially | Some can be deduced from text: most voted gueries |
| . 500100011 | or are experimental autosets) | . a. aay | Java projects with at least 20 stars |
| Analysis | Descriptive statistics | Yes | Boxplots for Frank and Precision@k (success-rate |
| • | • | | and MRR are averaged metrics) |
| | Inferential statistics | Yes | For Frank and precision@k (the others are averaged |
| | | | metrics) |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

Approach to debug method names based on the analysis of consistency between method names and method code AP3 $\,$ Paper

ID

Experiments 3 RQ4 is not an experiment but other type of empirical study

Comments Preprocessing is done with state-of-the-art NNs. There is a final step that is a regular algorithm. When compares agains other approaches, they do not c

| Aspect | Element | E1 | Comments |
|---------------------|---|---|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 7 (from Fig. 4) | Number of layers taken from figure |
| | | #neurons/layer: 1000 (?). Input layer is nxk k=94 | Table IV. Number of nodes in hidden layers is |
| | | connections: L2, L4 (convolutional), L3, L5 | 1000 but not clear if is total number or per |
| | | (subsampling, maxpool), L6 (dense) | layer(?). |
| | | activation functions: softmax (output), ReLU (rest) | |
| | | params. initialization: No | |
| | Model parameters | biases: No | |
| | · | weights: No | |
| | DL algorithm | representation: Yes | |
| | 3.0 | model type (architecture): Yes | |
| | | loss function: mean absolute error | |
| | | regularization: No | |
| | | optimization: SGD | |
| | Training hyperparameters | train-test split: Different datasets | Refers to [43] and [60] for parameters used |
| | Training Hyperparameters | learning rate: 1e-2 | nerels to [45] and [66] for parameters asea |
| | | #iterations: No | |
| | | batch size: No | |
| | | | |
| | Training data | #epochs: No | Described in paper and included in artifact |
| 0 | Training data Factors and treatments | Yes Partially | Described in paper and included in artifact DNN proposed is assessed, but not compared |
| | Response variable, elaboration and metric | Precision, recall, F1, accuracy | All of them perfectly defined |
| Design | Design type | No | |
| | Diaglina variables | No | |
| | Blocking variables Held-constant variables | No No | |
| | Measured variables (covariates) | Yes | K (size of sets of adjacent vectors). Outside of |
| | ivicasureu variabies (covariates) | 163 | DNN |
| | Randomization | No | DIVIN |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Yes | Built separately from training |
| mot different con | Measuring instruments | No | Same Separatery from daming |
| | Measurement procedure | No | |
| | Technological infrastructure | DL4J library | Implements using paragraph vector, wrod2vec |
| | 0 | - , | and LeNet5 (which does seem to lack a dense |
| | | | layer, but it is not clear what the real |
| | | | architecture is) |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| , | Inferential statistics | No | |
| Validity evaluation | | Internal, external | |
| Artifact | Availability | Github | Updated recently!!! (8 months ago) |
| Aidiact | | | |

Approach to debug method names based on th $\ensuremath{\mathsf{AP3}}$ Paper

ID. Experiments 3

Comments Preprocessing is done with state-of-the-art NNslo exactly the same task $% \left\{ 1,2,\ldots ,n\right\} =0$

| Aspect | Element | E2 | Comments |
|---------------------|--|--|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hypomacometers | Same as provious | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | DNN proposed is assessed, but not compared |
| Design | Response variable, elaboration and metric Design type | Inconsistency avoidance (T1), first-token accuracy (T2), full-name accuracy (T3) | Fully defined. Should be a nested design. |
| 203.6.1 | 200.8 1, pc | | Not clear if RV could be a factor |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | Yes | K (size of sets of adjacent vectors) and R (ranking strategy). Outside of DNN |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Same as previous | |
| | Measuring instruments | No | |
| | Measurement procedure | No Sama an arravious | |
| | Technological infrastructure | Same as previous | |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Same as previous | |
| | | | |
| Artifact | Availability | Same as previous | |

Approach to debug method names based on th $\ensuremath{\mathsf{AP3}}$ Paper

ID Experiments 3

Comments Preprocessing is done with state-of-the-art NNs

| Aspect | Element | E3 | Comments |
|---------------------|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | Nothing is said about the state-of-the-art DNN, CAN. Apparently the CAN model has been made available by the authors. They do not mention any kind of change to the original proposal. Just that the same training set is used for CAN and their proposal |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Model (n-gram, CAN:conv_attention, CAN:copy_attention, proposal:R1, proposal:R2, proposal:R3, proposal:R4) | n-gram is not a DNN. This time seems easier to identify |
| | Response variable, elaboration and metric | Precision, recall, F1 T1, T2, T3 | All of them perfectly defined |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | Yes | threshold |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Same as previous | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| Analysis | Descriptive statistics | No | |
| , | Inferential statistics | No | |
| Validity evaluation | | Same as previous | |
| Artifact | Availability | Same as previous | _ |
| | Badge | Same as previous | |
| | | Tama do proviodo | |

Paper On learning meaningufl code changes via neural machine translation

ID AP4
Experiments 2

Comments Optimization are hidden. Impossible to know how many

| Aspect | Element | E1 | Comments |
|---------------------|---|------------------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | same as next, but layers and | |
| | 7,000 | units are factors | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as next | |
| | Woder parameters | Same as next | |
| | DI alaasishaa | Same as next | |
| | DL algorithm | same as next | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as next | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Available in artifact |
| Operationalization | Factors and treatments | Yes | Type of RNN Cell (LSTM, GRU), number of |
| | | | layers for the encoder/decoder (1,2,4), |
| | | | number of units for the encoder/decoder |
| | | | (256,512), embedding size (256, 512) |
| | Response variable, elaboration and metric | Partially | Loss function, but we do not know which one |
| | , | , | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Available in artifact |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Partially | Some characteristics can be deduced from |
| | • | • | text |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing conclusion, but no statistical tests |
| , | , , | , | used |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | 20050 | | |

Paper On learning meaningufl code changes

ID AP4 Experiments 2

Comments Optimization are hidden. Impossible t

| Aspect | Element | E2 | Comments |
|---|---|--|--|
| Experiment type | | Evaluation | Assessment, does not compare |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | The paper does not mention which of the models |
| | | #neurons/layer: No | evaluated in E1 is finally chosen |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | Dealgonain | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | · · | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 60k | |
| | | | Available in artifact |
| | Training data | Yes | Available ili artifact |
| Operationalization | Training data Factors and treatments | Test set | Model is only assessed |
| Operationalization | | | Model is only assessed Raw count and percentage of successfully |
| | Factors and treatments Response variable, elaboration and metric | Test set Yes | Model is only assessed |
| Operationalization Design | Factors and treatments Response variable, elaboration and metric Design type | Test set Yes No | Model is only assessed Raw count and percentage of successfully |
| | Factors and treatments Response variable, elaboration and metric Design type Blocking variables | Test set Yes No No | Model is only assessed Raw count and percentage of successfully |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables | Test set Yes No No No | Model is only assessed Raw count and percentage of successfully |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) | Test set Yes No No No No No | Model is only assessed Raw count and percentage of successfully |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | Test set Yes No No No No No No | Model is only assessed Raw count and percentage of successfully |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | Test set Yes No No No No No No No No No | Model is only assessed Raw count and percentage of successfully |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run Available in artifact |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measurements Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Test set Yes No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run Available in artifact |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | Test set Yes No Partially | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run Available in artifact |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Test set Yes No Partially No | Model is only assessed Raw count and percentage of successfully predicted code changes Seems 1 run Available in artifact |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Test set Yes No Partially No No | Raw count and percentage of successfully predicted code changes Seems 1 run Available in artifact Some characteristics can be deduced from text |

Paper NL2Type: Inferring JavaScript function types from natural language information

ID AP5 Experiments 4

Comments 2 non-comparative experiments (RQ1 and RQ5) and a qualitative (RQ3)

| Aspect | Element | E1 | Comments |
|---------------------|---|---|---|
| Experiment type | | Optimization+Evaluation | Not sure what it is |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ5 |
| Variables selection | Statistical hypotheses Model hyperparameters | No Same as next | |
| valiables selection | wodernyperparameters | Jame as next | |
| | Model parameters | Same as next | |
| | DL algorithm | Same as next | |
| | Training hyperparameters | Same as next | |
| | Training data | Yes | Linked to artifact |
| Operationalization | Factors and treatments | DNN Model | NL2Type, NL2Type w/o comments, naive (always same answer, k most common types) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 on top-1-3-5 predicted Efficiency (average time perfunction or total) for NL2Type |
| Design | Design type | No | |
| | Blocking variables Held-constant variables | No No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration Procedure | No No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked to artifact |
| | Measuring instruments Measurement procedure | No No | |
| | Technological infrastructure | Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and | |
| | | NVIDIA Tesla P100 GPU with 16GB of memory. | |
| Population | Objects (chars. of the experimental datasets) | No | Mentions JavaScript files/libraries |
| Analysis | Descriptive statistics Inferential statistics | No No | |
| Validity evaluation | | No No | |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available, reusable |
| | | | |

Paper NL2Type: Inferring JavaScript function

ID AP5 Experiments 4

Comments 2 non-comparative experiments (RQ1

| Aspect | Element | E2 | Comments |
|---------------------|--|--|--|
| Experiment type | | Evaluation | Seems uses the "optimized" version |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| Variables selection | Statistical hypotheses Model hyperparameters | No Same as next? | |
| | Model parameters | Same as next? | |
| | DL algorithm | Same as next? | |
| | Training hyperparameters | Same as next? | |
| Operationalization | Training data Factors and treatments | Yes DNN Model | Linked to artifact DeepTyper (JSNice is not a DNN), NL2Type |
| | Response variable, elaboration and metric | Yes | For DeepTyper, use their publicly available artifact, and do not apply confidence threshold. Same test and training sets are used for both approaches Precision, recall, F1 on top-1 predicted |
| Desien | Designations | No | |
| Design | Design type Blocking variables Held-constant variables | No No | |
| | Measured variables (covariates) Randomization Task duration Procedure | No No No Partially | Create a front-end forNL2Type to use |
| | | | dataset in DeepTyper and allow fair comparison |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set Measuring instruments Measurement procedure | Yes No No | Linked to artifact |
| | Technological infrastructure | Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory. | |
| Population | Objects (chars. of the experimental datasets) | No | Mentions JavaScript files/libraries |
| Analysis | Descriptive statistics Inferential statistics | No No | , |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability Badge | Yes Yes | Available, reusable |
| | | | |

Paper NL2Type: Inferring JavaScript function

ID AP5 Experiments 4

Comments 2 non-comparative experiments (RQ1

| Aspect | Element | E3 | Comments |
|------------------------|--|--|---|
| Experiment type | | Evaluation | Evaluate DNN for another task (inconsistencies |
| _ | | | detection) |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| Variables selection | Statistical hypotheses Model hyperparameters | No Same as next? | |
| | , p, p | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as next? | |
| | | | |
| | DL algorithm | Same as next? | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as next? | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Linked to artifact |
| Operationalization | Factors and treatments | DNN | 1 treatment only |
| | | | |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Yes | Frequency of potential inconsistency types |
| | nesponse variable, elaboration and metric | 163 | (inconsistency/non-standard type |
| | | | annotation/misclassification) |
| Design | Design type | No | |
| | Blocking variables | No. | |
| | Held-constant variables | Partially | |
| | | | |
| | | | |
| | | | |
| | Measured variables (covariates) | No | |
| | Randomization Task duration | No No | |
| | Procedure | Partially | Multiple runs to check the predictions. Some neurons |
| | | | are purposefully deactivated during prediction. |
| | | | |
| | Number of experimental units | No | Seems more than 1 |
| Instrumentation | Test set Measuring instruments | Yes Yes | Linked to artifact |
| | Measurement procedure | Yes | NL2Type is used in a different way. The return value is |
| | | | used to check if the predicted type matches the real |
| | | | one. |
| | Technological infrastructure | Implemented in Python | |
| | | Preprocessing: Python NLTK library Word2Vec | |
| | | DNN: Keras | |
| | | Ubuntu 16.04 computer with Intel Xeon | |
| | | E5-2650 with 48 cores, 64GB memory | |
| | | and NVIDIA Tesla P100 GPU with 16GB | |
| | | of memory. | |
| | | | |
| Donulation | Objects (chars of the experience of the experien | No | Mantions JavaScript files /libraries |
| Population Analysis | Objects (chars. of the experimental datasets) Descriptive statistics | No No | Mentions JavaScript files/libraries |
| | Inferential statistics | No | |
| Validity evaluation | | No | |
| Artifact | Availability | Yes Yes | Available reusable |
| | Badge | 163 | Available, reusable |

Paper NL2Type: Inferring JavaScript function

ID AP5 Experiments 4

Comments 2 non-comparative experiments (RQ1

| Aspect | Element | E4 | Comments |
|---------------------|--|--|---|
| Experiment type | | Optimization | It should be the first experiment |
| Hypotheses | Research hypotheses Statistical hypotheses | Yes No | RQ4 |
| Variables selection | Model hyperparameters | #layers: 4 #neurons/layer: I1 (?), I2 (100?), I3 (256?), I4 (1000?) connections: input, bi-directional LSTM, fully connected, output (softmax) activation functions: No params. Initialization: No | Embedding built upon Word2Vec, apparently being retrained, but not clear if it is re-trained separately or together with DNN proposed. I have excluded from the paper the model hyperparameters of Word2Vec, as its architecture is not described (word embedding size: 100, context size: 5, min. ocurrence of word: 5) |
| | Model parameters | biases: No | #neurons deduced But they are in the artefact. |
| | DL algorithm | weights: No representation: Yes model type: Yes loss function: categorical cross entropy regularization: dropout (20%) optimization: Adam (defaults?) | |
| | Training hyperparameters | train-test split: 80-20 learning rate: No #iterations: No batch size: 256 | No need of K-cross validation, due to large amount of data |
| | Training data | #epochs: 12 Yes | Linked to artifact |
| Operationalization | Factors and treatments | DNN architecture | output of DNN (55000) Paper mentions they have run experiments to choose hyperparameters, but they are not described |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 on top-1 predicted |
| Design | Design type | No | |
| | Blocking variables Held-constant variables | No Partially | Input representation: words in names: 6 words in comment: 12 words in comment: 10 #pars: 10 |
| | Measured variables (covariates) | No | pai.5. 20 |
| | Randomization | No | |
| | Task duration Procedure | No No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Linked to artifact |
| | Measuring instruments Measurement procedure | No No | |
| | Technological infrastructure | Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory. | |
| Population | Objects (chars. of the experimental datasets) | No | Mentions JavaScript files/libraries |
| Analysis | Descriptive statistics | No | |
| Validity evaluation | Inferential statistics Conclusion, internal, construct, external | No No | |
| | | Yes | |
| Artifact | Availability | 163 | |

Paper ActionNet: Vision-based workflow action recognition from programming screencasts

ID AP6
Experiments 5

Comments One more study (with no RQ) but it is not an experiment

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|--|
| Experiment type | | Optimization | |
| | | | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: CNNxn+softmax | They propose 2 different architectures. Early fusion |
| | | #neurons/layer: No | vs late fusion. But layers are the same. Late fusion |
| | | connections: No | is siamese. Probably #neurons/layer would |
| | | activation functions: No | change. |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Partially | |
| | 3 | model type: Yes | |
| | | loss function: Yes | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | 80%-20% |
| | Training hyperparameters | learning rate: No | 3070-2070 |
| | | #iterations: No | |
| | | | |
| | | batch size: No | |
| | | #epochs: No | |
| Operationalization | Training data Factors and treatments | Yes Yes | 80% |
| Operationalization | ractors and treatments | 165 | Input change region strategies (change-contrast, |
| | | | action-continuity, both) |
| | | | DNN architecture (early vs late fusion) |
| 5 . | Response variable, elaboration and metric | Yes | Accuracy, precision, recall, F1 |
| Design | Design type | Partially | 2-factor (deduced, as mentions 6 models) |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) Randomization | No | |
| | | No No | |
| | Task duration Procedure | No Partially | Each model is trained with same data |
| | Number of experimental units | No | Laci moderis trained with same data |
| Instrumentation | Test set | Yes | 20% |
| mod dimentation | Measuring instruments | No | 2070 |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | It is described in a non-comparative experiment |
| | | , | (RQ3) |
| | | | PC with 64GB RAM, i9-7900x CPU, Titan Xp GPU |
| | | | (missing OS) |
| Population | Objects (chars. of the experimental datasets) | Yes | Python and Java. Also describe the procedure |
| . opulation | o operation of the experimental additions | | followed |
| Analysis | Descriptive statistics | Partially | Mean per strategy and architecture |
| Allalysis | Inferential statistics | No | Weari per strategy and architecture |
| Validity evaluation | | No | |
| Artifact | Availability | No | |
| | Badge | No | |
| | 50050 | | |

Paper ActionNet: Vision-based workflow act

ID AP6 Experiments 5

Comments One more study (with no RQ) but it is

| Aspect | Element | E2-E4 | Comments |
|---------------------|---|-------------------------------|--|
| xperiment type | | Generalization | Are 3 experiments. Behaviour in other |
| | | | developers/working environmnets/programming |
| | | | languages |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| /ariables selection | Model hyperparameters | Best from factors in previous | both and early-fusion |
| | | experiment | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | Train-test split: |
| | | | - 80%-20% for intraplaylist experiment |
| | | | - 4-1 for interplaylist experiment |
| | | | - 5-5 for inter-language |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | 1 treatment DNN |
| | Response variable, elaboration and metric | Yes | Accuracy, precision, recall, F1 |
| Design | Design type | Partially | Deduced |
| | Blocking variables | No | Playlist in 2 experiments, language in 1??? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | Partially | |
| | Number of experimental units | No | |
| nstrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | It is described in a non-comparative experiment (RQ3) |
| | | | PC with 64GB RAM, i9-7900x CPU, Titan Xp GPU (missing |
| | | | OS) |
| Population | Objects (chars. of the experimental datasets) | Yes | Python and Java. Also describe the procedure followed |
| Analysis | Descriptive statistics | Partially | Mean and stddev per playlist (E2, E3) or action class (E4) |
| | Inferential statistics | No | |
| /alidity evaluation | Conclusion, internal, construct, external | No | |
| | Availability | No | |
| Artifact | , transaction, | | |

Paper ActionNet: Vision-based workflow act

ID AP6
Experiments 5

Comments One more study (with no RQ) but it is

| Aspect | Element | E5 | Comments |
|---------------------|--|---------------------|---|
| Experiment type | | Evaluation | Time |
| | | | Probably time is measured in E1 and this would not be a |
| | | | new experiment. But nothing is said |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| Try potricses | Statistical hypotheses | No | ngs |
| Variables selection | Model hyperparameters | No | late-fusion. Does not mention anything else |
| variables selection | modern) perparameters | | ate issue sees itemental any amy esse |
| | | | |
| | Model parameters | No | |
| | DL algorithm | No | |
| | | | |
| | Training hyperparameters | No | |
| | | | |
| | | | |
| Operationalization | Training data Factors and treatments | No Yes | 1 treatment DNN |
| Operationalization | ractors and treatments | res | I treatment DNN |
| | | | |
| | December of the alphanetics and section | Dente III. | T: |
| Design | Response variable, elaboration and metric Design type | Partially Partially | Time Deduced |
| Design | Blocking variables | No | Deduced |
| | Held-constant variables | No | |
| | | | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | It is described in a non-comparative experiment (RQ3) |
| | | , | PC with 64GB RAM, i9-7900x CPU, Titan Xp GPU (missing |
| | | | |
| | | | OS) |
| Population | Objects (chars. of the experimental datasets) | No | Python and Java. Also describe the procedure followed |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| | | ** | |
| Artifact | Availability | No | |

Paper Training binary classifiers as data structure invariants

ID AP7
Experiments 3

Comments RQ1 experiments with dataset generation, not DNN

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|--|
| Experiment type | | Optimization | Random search used |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 3 (input, hidden, | |
| | | output) | |
| | | #neurons/layer: L1: ?, L2: | |
| | | factor, L3: 1 | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: factor | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Different sets are used |
| | Training Tryperparameters | learning rate: No | Sincrent sets are asea |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Same as next | |
| | Hallillig data | Sallie as liext | |
| Operationalization | Factors and treatments | Yes | Number of units in hidden layer (2,100), regularization (-5,3). 10 random combinations of them |
| | Response variable, elaboration and metric | No | Just mentions best performance (and reduce validation set error) |
| Design | Design type | No | |
| · · | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | No | |
| | | | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | scikit-learn |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

Paper Training binary classifiers as data stru

ID AP7 Experiments 3

Comments RQ1 experiments with dataset genera

| Aspect | Element | E2 | Comments | |
|---------------------|---|------------------|---|--|
| Experiment type | | Evaluation | | |
| Hypotheses | Research hypotheses | Yes | RQ2 | |
| | Statistical hypotheses | No | | |
| Variables selection | Model hyperparameters | Same as previous | | |
| | | | | |
| | | | | |
| | | | | |
| | Model parameters | Same as previous | | |
| | DL algorithm | Same as previous | | |
| | | | | |
| | | | | |
| | Training hyperparameters | Same as previous | | |
| | | | | |
| | | | | |
| | Training data | Partially | The sentence referring to the artifact is vey | |
| | | | generic (experiments can be reproduced | |
| | | | following the instructions found in the site of the RP of our approach), and we do not know if the | |
| | | | dataset is provided, or the steps to generate the | |
| | | | dataset | |
| Operationalization | Factors and treatments | Partially | Model type (Daikon, proposal) | |
| | | D 11 II | | |
| | Response variable, elaboration and metric | Partially | number of objects correctly and incorrectly classified, precision, recall (no explanation), | |
| | | | training time | |
| Design | Design type | No | | |
| | Blocking variables | Deduced | Instance (positive/negativa), scope | |
| | Held-constant variables | No | | |
| | Measured variables (covariates) | No | | |
| | Randomization | No | | |
| | Task duration | No | | |
| | Procedure | No | | |
| | Number of experimental units | No | | |
| Instrumentation | Test set | Partially | Same situation as training | |
| | | | Some instances of the test set might have appeared in the training set | |
| | Measuring instruments | No | | |
| | Measurement procedure | Yes | | |
| | Technological infrastructure | Partially | Same as previous | |
| Population | Objects (chars. of the experimental datasets) | No | | |
| Analysis | Descriptive statistics | No | _ | |
| | Inferential statistics | No | | |
| Validity evaluation | | Partially | _ | |
| Artifact | Availability | Yes | | |
| | Badge | No | | |

Paper Training binary classifiers as data stru

ID AP7
Experiments 3

Comments RQ1 experiments with dataset genera

| Aspect | Element | E3 | Comments | |
|---------------------|-------------------------|------------------|---------------------|--|
| Experiment type | | Generalization | Embedded in Randoop | |
| Hypotheses | Research hypotheses | Yes | RQ3 | |
| | Statistical hypotheses | No | | |
| Variables selection | n Model hyperparameters | Same as previous | | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

Training data Same as previous

| Operationalization | Factors and treatments | Partially | Model type (Daikon, proposal) embedded in Randoo |
|---------------------|---|-----------|---|
| | Response variable, elaboration and metric | Yes | Number of bugs found |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | Partially | This time scope=5 |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | Yes | For defect detection a timeout of 10 minutes is set |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Again, they merely describe where it is taken from |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Same as previous |
| Population | Objects (chars. of the experimental datasets) | No | <u> </u> |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | |
| Artifact | Availability | Yes | |
| | Badge | No | |

A novel neural source code representation based on abstract syntax tree AP8 $\,\,$ 4

Paper ID Experiments

| Aspect | Element | E1 | Comments |
|---------------------|--|----------------------------------|---|
| Experiment type | | Evaluation | Source code classification |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| (1d-lalala-st) | Statistical hypotheses | No CO | F |
| variables selection | Model hyperparameters | #layers: 6? | For encodder mentions Word2Vec. Its role not explained. |
| | | #neurons/layer: GRU (100) | Could be the pre-trained encoder (mesning weights are |
| | | connections: encoder, recurrent, | initialized with these values) |
| | | pooling, output | |
| | | activation functions: Some | |
| | | mentioned: identity (encoder) | |
| | | params. Initialization: No | |
| | Model parameters | biases: Yes | Explicitly says that "trained models are stored" |
| | | weights: Yes | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: cross-entropy | |
| | | regularization: No | |
| | | optimization: AdaMax | |
| | Training hyperparameters | train-test split: 60-20-20 | |
| | · | learning rate: 0.002 | |
| | | #iterations: No | |
| | | batch size: 64 | |
| | | #epochs: max. 15 | |
| | Training data | Yes | OJ. Referenced |
| perationalization | Factors and treatments | Partially | ASTNN, TextCNN, LSTM, LSCNN |
| , | | , | For other approaches: |
| | | | TextCNN: kernel size=3, filters=100 |
| | | | LSTM: hidden states =100 |
| | | | LSCNN: nothing |
| | Response variable, elaboration and metric | Yes | Accuracy |
| esign | Design type | No | Seems 1 factor-6 treatment (TextCNN, LSTM, |
| 6 | 81/6- | | TBCNN,LSCNN,PGD+GGNN) |
| | Blocking variables | No | 150111,250111,105.001111) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | | |
| | Number of experimental units | No | Assume 1 |
| strumentation | Test set | Yes | OJ. Referenced |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | Partially | pycparser (C) and javalang (Java) to obtain ASTs |
| | | | train embeddings using word2vec (embedding size=128) |
| | | | 16 cores of 2.4GHz CPU, Titan Xp GPU |
| opulation | Objects (chars. of the experimental dataset) | Partially | Mention the datasets and references (OJ) |
| nalysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partially | Only 3 threats are listed, not classified |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |

A novel neural source code represent AP8 4

Paper ID Experiments Comments

| Aspect | Element | E2 | Comments |
|---------------------|------------------------|------------------|----------------------|
| Experiment type | | Evaluation | Code clone detection |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous Loss function is binary cross-entropy

Same as previous #epochs: max. 5 Threshold: 0.5 Training hyperparameters

| | Training data | Yes | OJ, BCB, referenced |
|--------------------|---|------------------|--|
| Operationalization | Factors and treatments | Partially | ASTNN, RAE+, CDLH |
| | | | For other approahces: |
| | | | RAE+: Configuration as in paper |
| | | | CDLH: Not public, results from paper |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | Seems 1 factor-4 treatment (RAE+,CDLH,PGD+GGNN |
| | | | ASTNN) |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Assume 1 |
| Instrumentation | Test set | Yes | OJ, BCB, referenced |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | Same as previous | |

| Objects (chars. of the experimental dataset) | Partially | Mention the datasets and references (OJ, BCB) |
|--|--|--|
| | | OJ seems to be different from the one used in E1 |
| Descriptive statistics | No | |
| Inferential statistics | No | |
| Conclusion, internal, construct, external | Partially | Only 3 threats are listed, not classified |
| Availability | Yes | |
| Badge | Yes | Available badge |
| | Descriptive statistics Inferential statistics Conclusion, internal, construct, external Availability | Descriptive statistics No Inferential statistics No Conclusion, internal, construct, external Partially Availability Yes |

Paper ID A novel neural source code represent AP8 4

Experiments

| Aspect | Element | E3 | Comments |
|---------------------|------------------------|------------------|-------------------------------|
| Experiment type | | Optimization | Several architectural choices |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hypernarameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Yes | OJ, BCB, referenced |
|---------------------|--|------------------|---|
| Operationalization | Factors and treatments | Partially | AST-full/block/node |
| | | | Removing pooling I/II |
| | | | LSMT instead of GRU |
| | | | long code fragments |
| | | | ASTNN |
| | Response variable, elaboration and metric | Yes | Accuracy, F1 |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Assume 1 |
| Instrumentation | Test set | Yes | OJ, BCB, referenced |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | Same as previous | |
| Population | Objects (chars. of the experimental dataset) | Same as previous | |
| | | | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | | Partially | Only 3 threats are listed, not classified |
| Artifact | Availability | Yes | |

A novel neural source code represent AP8 4

Paper ID Experiments

| Aspect | Element | E4 | Comments |
|---------------------|------------------------|------------------|--------------------|
| Experiment type | | Optimization | Batching algorithm |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | No | Not clear which ones are used |
|---------------------|--|------------------|---|
| Operationalization | Factors and treatments | Partially | without batching |
| | | | batching recurrent layer |
| | | | batching recurrent+enconding layers |
| | Response variable, elaboration and metric | Yes | Time |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Assume 1 |
| Instrumentation | Test set | No | Not clear which ones are used |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | Same as previous | |
| Population | Objects (chars. of the experimental dataset) | No | Not clear the ones used |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Only 3 threats are listed, not classified |
| Artifact | Availability | Yes | • |
| | Badge | Yes | Available badge |

Paper A neural model for generating natural language summaries of program subroutines

Paper A ne
ID AP9
Experiments 2
Comments

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | |
| | Statistical hypotheses | No | 51 |
| Variables selection | Model hyperparameters | #layers: Yes | 5 layers: input (x3), Embedding (x3), GRU(x3), |
| | | #neurons/layer: Partially | Attention (x2), Dense (1) |
| | | connections: Yes | Input size: 100, 123, 100 |
| | | activation functions: No | Embedding size: ? |
| | | params. Initialization: No | GRU sizes: 256 each |
| | | | All this is deduced from text. Other numbers are |
| | | | given, but it is not straightforward to see what |
| | Model parameters | biases: No | they are |
| | Woder parameters | weights: No | |
| | DL algorithm | representation: Yes | The missing info can be found in the code. |
| | DE algoritation | model type: Yes | This is a good example of code and paper being |
| | | loss function: No | inconsistent!!!! |
| | | regularization: No | inconsistent:::: |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | 90%-5%-5% |
| | Training hyperparameters | learning rate: | All DNNs 10 epochs |
| | | #iterations: | All DIVING TO Epochs |
| | | batch size: | |
| | | #epochs: 10 | |
| | Training data | Yes | 95% dataset. |
| Operationalization | Factors and treatments | Partially | DNN: attendgru, SBT, codenn, ast-attendgru |
| орегинопинации | . actors and deadnesses | r dradiny | Mention that for codenn they use their publicly |
| | | | available implementation given its complexity. |
| | | | Attendgru seems their approach without the AST |
| | | | encoder. |
| | | | SBT seems to have been modified |
| | | | Nothing else is said about them |
| | Response variable, elaboration and metric | Partially | BLEU14, composite BLEU (?) |
| | | , | Formulas are not given |
| Design | Design type | No | |
| 0 | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Just one? |
| Instrumentation | Test set | Yes | 5% dataset |
| | Measuring instruments Measurement procedure | No No | |
| | Technological infrastructure | Partially | Xeon E5-1650v4 CPU, 64GB RAM, 2 Quadro |
| | recimological illitastructure | raradily | P5000 GPUs. GPUs with 16GB VRAM were |
| | | | necessary due to the large size of the model. |
| | | | Missing OS. Implemented with Keras. |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Partially | Java methods from the Sourcerer repository |
| | = 1,111 (1.m.s. o. a.e experimental datasets) | | provided by Lopes et al. |
| Analysis | Descriptive statistics | No | F 1 |
| , | Inferential statistics | No | |
| Validity evaluation | | Partially | Present a list of 2 threats. No division |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

A neural model for generating natural

ID AP9 2 Experiments Comments

| Aspect | Element |
|---------------------|------------------------|
| Experiment type | |
| Hypotheses | Research hypotheses |
| | Statistical hypotheses |
| Variables selection | Madalhunamaramatara |

| E2 | Comments |
|------------------|---|
| Optimization | Assuming only AST available (no internal documentation) |
| Yes | |
| No | |
| Same as previous | Only difference is dataset used |

| Experiment type | | Optimization | Assuming only AST available (no internal docu |
|--------------------|--|-------------------|---|
| Hypotheses | Research hypotheses | Yes | |
| | Statistical hypotheses | No | |
| ariables selection | Model hyperparameters | Same as previous | Only difference is dataset used |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | Training data | Same as previous | Code removed |
| Operationalization | Factors and treatments | Partially | ast-attendgru without source code |
| | Secretary of the state of the s | Position . | DIFINAL Accessión DIFINA |
| | Response variable, elaboration and metric | Partially | BLEU14, composite BLEU (?) Formulas are not given |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Just one? |
| nstrumentation | Test set | Same as previous? | Code removed |
| | Measuring instruments | No | |
| | Measurement procedure Technological infrastructure | No Partially | Xeon E5-1650v4 CPU, 64GB RAM, 2 Quadro |
| | reciniologicarinirastructure | raitially | P5000 GPUs. GPUs with 16GB VRAM were |
| | | | necessary due to the large size of the model. Missing OS. Implemented with Keras. |
| Population | Objects (chars. of the experimental datasets) | Partially | Java methods from the Sourcerer repository provided by Lopes et al. |
| Analysis | Descriptive statistics | No | |

| mstumentation | Measuring instruments Measurement procedure Technological infrastructure |
|---------------------------------|--|
| Population | Objects (chars. of the experimental datasets) |
| Analysis | Descriptive statistics Inferential statistics |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability Badge |

| No | |
|-------------------|--|
| No | |
| No | Just one? |
| Same as previous? | Code removed |
| No | |
| No | |
| Partially | Xeon E5-1650v4 CPU, 64GB RAM, 2 Quadro |
| | P5000 GPUs. GPUs with 16GB VRAM were |
| | necessary due to the large size of the |
| | model. Missing OS. Implemented with |
| | Keras. |
| Partially | Java methods from the Sourcerer repository |
| | provided by Lopes et al. |
| No | |
| No | |
| Partially | Present a list of 2 threats. No division |
| Yes | |
| No | |

DeepPerf: performance prediction for configurable software with deep sparse neural network AP10 4 Paper

Experiments

E1 are a series of experiments. Difficult to assess how many, as they are described at a very high level

| Aspect | Element | E1 | Comments |
|---------------------------------|---|--|---|
| Experiment type | | Optimization | Hyperparameters tuning. Could be several experiments. Very bad described |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: n+2 | |
| | | #neurons/layer: I1 (), I2n+1(128?), | |
| | | In+2(1) | |
| | | connections: | |
| | | activation functions: | |
| | | params. Initialization: Xavier (weights) | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | . 0 | model type: Yes | |
| | | loss function: Yes (MSE) | |
| | | regularization: L1 (only in l2). Lambda, | |
| | | grid search with 30 points | |
| | | logarithmically spaced in 0.01-1000 | |
| | | optimization: Adam (default Tensorflow | |
| | | values), gradient clipping | |
| | Training hyperparameters | train-test split: | |
| | rraining hyperparameters | learning rate: initial between 0.0001- | |
| | | 0.1, dropped by 0.001 | |
| | | #iterations: | |
| | | | |
| | | batch size: Size of training data | |
| | Total or date | #epochs: 2000? Yes | 1111111111 |
| | Training data | Yes | Input and output are normalized (0-1 and 0-100). |
| 0 | P. de and the state of the | Design III | Explicitly linked to artifact in paper |
| Operationalization | Factors and treatments | Partially | Could be several experiments. Not sure if all |
| | | | hyperparams are made explicit |
| | | | Factors (at least): regularization, #hidden layers, learning |
| | | | rate |
| | | | No levels given |
| | Response variable, elaboration and metric | No | time (?) |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | Partially | #neurons/layer, #epochs, but no value given |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration Procedure | No No | |
| | | No | Coome 1 run |
| lasta un catatian | Number of experimental units | | Seems 1 run |
| Instrumentation | Test set | Yes | Input normalized (0-1). Explicitly linked to artifact in |
| | Manager in the contract of | Ne | paper |
| | Measuring instruments | No | |
| | Measurement procedure | No Puthon 3.6. Tonsorflow 1.8.0 | |
| Population | Technological infrastructure Objects (chars. of the experimental datasets) | Python 3.6, Tensorflow 1.8.0 Partially | Briefly describes them. References are given to other |
| гориации | Objects (chars, or the experimental datasets) | raitiany | |
| Analysis | Descriptive statistics | No | publications where they are fully explained |
| Milaly SIS | * | No | |
| Validity ovaluation | Inferential statistics Conclusion, internal, construct, external | | Internal external |
| Validity evaluation Artifact | | Partially Yes | Internal, external |
| Aruiact | Availability | Yes Yes | Available |
| | Badge | 103 | Available |

DeepPerf: performance prediction for AP10

Paper ID

Experiments

E1 are a series of experiments. Difficu

| Aspect | Element | E2 | Comments |
|---------------------|---|-------------------------------------|--|
| xperiment type | | Evaluation | |
| | | | |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ4 |
| | Statistical hypotheses | No | |
| ariables selection | Model hyperparameters | Best from previous | |
| | | | |
| | | | |
| | Model parameters | Best from previous | |
| | DL algorithm | Best from previous | |
| | | | |
| | Training hyperparameters | Best from previous | |
| Operationalization | Training data Factors and treatments | Yes Model Type, Subject system (?) | Input and output are normalized (0-1 and 0-100) DECART, DeepPerf (DECART is classification trees) Not sure if subject system (apache, x264,8DB-J, LLVM, BDB-C, SQLite). Could be blocking variable |
| | Response variable, elaboration and metric | Yes | Mean Relative Error (MRE), training time |
| esign | Design type | No | |
| | Blocking variables | Yes | N-fold validation (30 times resampling training set) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | C1 |
| nctrum ontation | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | Yes | Input normalized (0-1). Explicitly linked to artifact in |
| | Measuring instruments | No | paper |
| | Measurement procedure | No | |
| | Technological infrastructure | Python 3.6, Tensorflow 1.8.0 | |
| opulation | Objects (chars. of the experimental datasets) | Partially | Briefly describes them. References are given to other |
| | | | publications where they are fully explained |
| Analysis | Descriptive statistics | Yes | Mean and 95% CI |
| | Inferential statistics | Yes | t-test |
| /alidity evaluation | | Partially | Internal, external |
| Artifact | Availability Badge | Yes | Available |
| | | Yes | |

DeepPerf: performance prediction for AP10

Paper ID Experiments

E1 are a series of experiments. Difficu

| Aspect | Element | E3 | Comments |
|---|--|---|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2, RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Best from previous | |
| | Model parameters | Best from previous | |
| | DL algorithm | Best from previous | |
| | Training hyperparameters | Best from previous | |
| Operationalization | Training data Factors and treatments | Yes DNN architecture | Input and output are normalized (0-1 and 0-100) SPLConqueror (no DNN), DeepPerf (DECART is classification trees) |
| | Response variable, elaboration and metric | Yes | Mean Relative Error (MRE), training time |
| Design | Design type Blocking variables | No Yes | N-fold validation (30 times resampling training set) |
| | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No No No No | Seems 1 run |
| nstrumentation | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No No No No No Yes | Seems 1 run Input normalized (0-1). Explicitly linked to artifact in paper |
| | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No No No No Yes No No Python 3.6, Tensorflow 1.8.0 | Input normalized (0-1). Explicitly linked to artifact in paper |
| | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No | Input normalized (0-1). Explicitly linked to artifact in paper Briefly describes them. References are given to other |
| Population | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No Yes No Python 3.6, Tensorflow 1.8.0 Partially Yes | Input normalized (0-1). Explicitly linked to artifact in paper |
| Population Analysis | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No No No Yes No Python 3.6, Tensorflow 1.8.0 Partially Yes No | Input normalized (0-1). Explicitly linked to artifact in paper Briefly describes them. References are given to other publications where they are fully explained Mean and 95% CI |
| nstrumentation Population Analysis Validity evaluation Artifact | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No Yes No Python 3.6, Tensorflow 1.8.0 Partially Yes | Input normalized (0-1). Explicitly linked to artifact in paper Briefly describes them. References are given to other publications where they are fully explained |

DeepPerf: performance prediction for AP10 4

Paper ID Experiments

E1 are a series of experiments. Difficu

Availability Badge

Artifact

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------------------|---|
| Experiment type | | Optimization | Different architectures (SVM, dropout, L1, L2, no |
| | | | regularization) |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Best from previous | |
| | | | |
| | Model parameters | Best from previous | |
| | DL algorithm | Best from previous | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Best from previous | |
| | | | |
| | | | |
| | Training data | Yes | Input and output are normalized (0-1 and 0-100) |
| Operationalization | Factors and treatments | DNN architecture | DeepPerf, L1-all-FNN, Plain-FNN, L2-FNN, Dropout-FNN Also SVM, but it is not a DNN |
| | | | For the others mentions some hyperparameters, but |
| | | | they are not fully described |
| | Response variable, elaboration and metric | Yes | Mean Relative Error (MRE), training time |
| Design | Design type | No | N feld of Paletter (20 Page 1997) |
| | Blocking variables | Yes | N-fold validation (30 times resampling training set) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization Task duration | No | |
| | Procedure | No No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Input normalized (0-1). Explicitly linked to artifact in |
| | | | paper |
| | Measuring instruments | No | ETET. |
| | Measurement procedure | No | |
| | Technological infrastructure | Python 3.6, Tensorflow 1.8.0 | |
| Population | Objects (chars. of the experimental datasets) | Partially | Briefly describes them. References are given to other |
| P | , | : <i>I</i> | publications where they are fully explained |
| Analysis | Descriptive statistics | Yes | Mean and 95%CI |
| | Inferential statistics | No | |
| Validity evaluation | | Partially | Internal, external |
| | | | |

Yes Yes

Available

Unblind your apps: Predicting natural-language labels for mobile gui components by deep learning

Paper ID AP11 2

| Experim | ents |
|---------|------|
| | |

| Aspect | Element | E1 | Comments |
|---------------------|--|---------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Difficult | Resnet-101 Architecture pre-trained on MS |
| | | #neurons/layer: No | COCO dataset as CNN module (removing |
| | | connections: Difficult | last pooling layer) |
| | | activation functions: Some | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | beta1, beta2 and epsilon given for Adam |
| | | model type: Yes | Loss function not clear if whole net or |
| | | loss function: Relative entropy | encoder/decoder only |
| | | regularization: No | • |
| | | optimization: Adam for | |
| | | encoder/decoder | |
| | Training hyperparameters | train-test split: Yes | 80% training, 10% test, 10% validation (for |
| | Training Hyperparameters | learning rate: Variable | each app. category) |
| | | #iterations:No | Formula for learning rate given |
| | | batch size:No | Torrida for rearring rate given |
| | | #epochs:No | |
| | Training data | Yes | Linked to github |
| Operationalization | Factors and treatments | Neural models | CNN+LSTM, CNN+CNN, LabelDroid |
| Operationalization | ractors and acadinents | rectial models | The other two approaches do not seem to |
| | | | be from SE proposals, but ideas from |
| | | | authors. Nothing is said about other |
| | | | approaches |
| | Decrees a sight a state and a serie | Domini | ••• |
| Design | Response variable, elaboration and metric Design type | Partial No | Missing description of metric |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Missing OS | Pytorch, Intel i7-7800X CPU, 64GB RAM, |
| | | | NVIDIA GeForce GTX 1080 Ti GPU |
| Population | Objects (chars. of the experimental datasets) | Yes | Can be downloaded |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | Yes | Not explained why non-parametrics |
| | | | Not clear which are the datapoints |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| | Availability | Yes | https://github.com/chenjshnn/LabelDroid |
| Artifact | Availability | | https://github.com/chemphini/cuberbroid |

Unblind your apps: Predicting natural-AP11 2

Paper ID Experiments

| Aspect | Element | E2 | Comments |
|---|--|---|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| ** | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| Operationalization | Factors and treatments | Approach | |
| Operationalization | | | |
| Operationalization | | | |
| | Factors and treatments | Approach | |
| | Factors and treatments Response variable, elaboration and metric Design type Blocking variables | Approach Yes | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables | Approach Yes No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) | Approach Yes No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | Approach Yes No No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | Approach Yes No No No No No No No No No | Looke like 4 min |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units | Approach Yes No No No No No No No No No | Looks like 1 run |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set | Approach Yes No No No No No No No Yes | |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | Approach Yes No No No No No No No No No | 5-point Likert scale |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | Approach Yes No No No No No No No No No Yes Yes Yes | |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | Approach Yes No No No No No No No Yes Yes Yes | 5-point Likert scale Expert assessment |
| Design Instrumentation | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | Approach Yes No No No No No No No No No Yes Yes Yes | 5-point Likert scale Expert assessment |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | Approach Yes No No No No No No No Yes Yes Yes Yes Missing OS | 5-point Likert scale Expert assessment |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Approach Yes No No No No No No No No No See Wes Yes Yes Yes Yes | 5-point Likert scale Expert assessment Same as before |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Approach Yes No No No No No No No No No So Ves Yes Yes Yes Yes Yes | 5-point Likert scale Expert assessment Same as before Mean and std (per app type and total) |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Approach Yes No No No No No No No No No So Ves Yes Yes Yes Yes Yes | 5-point Likert scale Expert assessment Same as before Mean and std (per app type and total) Seems incorrect Wilcoxon (should be Friedman) |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Approach Yes No No No No No No No No No So Ves Yes Yes Yes Yes Yes | 5-point Likert scale Expert assessment Same as before Mean and std (per app type and total) Seems incorrect Wilcoxon (should be Friedman) Not explained why non-parametrics |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Approach Yes No No No No No No No No Yes Yes Yes Yes Yes Yes Yes | 5-point Likert scale Expert assessment Same as before Mean and std (per app type and total) Seems incorrect Wilcoxon (should be Friedman) Not explained why non-parametrics |

Paper CC2Vec: Distributed representations of code changes

ID: AP12 Experiments 4

| Con | nm | er | its |
|-----|----|----|-----|

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|--|
| Experiment type | | Evaluacion | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: yes | The architecture is described only |
| | | #neurons/layer: No | SOTA approaches are not described |
| | | connections: partially | |
| | | activation functions: some | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | Not mentioned |
| | • | weights: No | |
| | DL algorithm | representation: Yes | DL algorithm is fully described. Are options |
| | 220,601.0 | model type: Yes | in regularization and optimization |
| | | loss function: Yes | described? |
| | | regularization: Yes | described: |
| | | • | |
| | | optimization: Adam | - ! 6! 1 1 |
| | Training hyperparameters | train-test split: No | They refer to the whole sets used, taken |
| | | learning rate: No | from previous papers. But nothing else is |
| | | #iterations: No | explained. |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Partially | Mentions whole dataset |
| Operationalization | Factors and treatments | Neural model: CC2Vec, NMET | |
| | Response variable, elaboration and metric | Yes | BLEU-4 |
| Design | Design type | No | |
| _ | Blocking variables | Yes | Datset could be (or maybe factor) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| nstrumentation | Test set | Partially | Mentions whole dataset |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| opulation | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| /alidity evaluation | Conclusion, internal, construct, external | Internal, external | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper CC2Vec: Distributed representations o

ID: AP12 Experiments 4

Comments

| Aspect | Element | E2 | Comments |
|---------------------------------|--|--------------------|---|
| Experiment type | | Generalization | CC2Vec used to complement SOTA approaches |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | SOTA approaches are not described |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | They refer to the whole sets used, taken from previous papers. But nothing else is explained. |
| | Training data | Partially | Mentions whole dataset |
| Operationalization | Factors and treatments | Neural model | CC2Vec used to improve SOTA (patchnet, and a non-NN approach) |
| | Response variable, elaboration and metric | Yes | Accuracy, precision, recall, F1, AUC |
| Design | Design type | No | |
| | Blocking variables | Yes | Fold-cross-validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run, but mention 5-fold cross-validation |
| Instrumentation | Test set | Partially | Mentions whole dataset |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | · | totament automoral | |
| Validity evaluation | Conclusion, internal, construct, external | Internal, external | |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Yes | |

Paper CC2Vec: Distributed representations o

ID: AP12 Experiments 4

Comments

| Aspect | Element | E3 | Comments |
|---------------------|---|-------------------------------|---|
| Experiment type | | Generalization | CC2Vec used to complement SOTA approaches |
| Hypotheses | Research hypotheses | No | - |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | SOTA approaches are not described |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | They refer to the whole sets used, taken from previous papers. But nothing else is explained. |
| Operationalization | Training data Factors and treatments | Same as previous Neural model | Mentions whole dataset CC2Vec used to improve SOTA DeepJIT |
| | | | |
| Di | Response variable, elaboration and metric | Yes | AUC |
| Design | Design type | No | Fold-cross-validation |
| | Blocking variables | Yes | FOIG-Cross-validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) Randomization | No | |
| | Task duration | No | |
| | | No | Looks like 1 run but montion E fold cross validation |
| Instrumentation | Number of experimental units Test set | No Partially | Looks like 1 run, but mention 5-fold cross-validation Mentions whole dataset |
| mstrumentation | Measuring instruments | No | ivientions whole dataset |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| Allulysis | Inferential statistics | No | |
| Validity evaluation | | Internal, external | |
| Artifact | Availability | Yes | |
| Artifact | Badge | No | |
| | Dauge | INO | |

Paper CC2Vec: Distributed representations o

ID: AP12 Experiments 4

Badge

Comments

| Aspect | Element | E4 | Comments |
|---------------------|---|---|--|
| Experiment type | | Optimization | Experiment run for the three previous datasets |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Parially Comparison function (7 levels) | The 3 datasets of previous experiments Different comparison functions in the comparison layer |
| | Decrease variable elaboration and matric | Vac | DIELLA E1 ALIC |
| Design | Response variable, elaboration and metric | Yes No | BLEU-4, F1, AUC |
| Design | Design type | | Datacat? |
| | Blocking variables | Yes | Dataset? |
| | Held-constant variables | No | |
| | Measured variables (covariates) Randomization | No | |
| | Task duration | No No | |
| | | | Nothing acoustinged your about F fold |
| Instrumentation | Number of experimental units Test set | No Partially | Nothing mentioned now about 5-fold |
| mstrumentation | | • | Same as previous |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| Donulation | Technological infrastructure | No Partially | Samo as provious |
| Population | Objects (chars. of the experimental datasets) | Partially | Same as previous |
| Analysis | Descriptive statistics | No | |
| Maltalia | Inferential statistics | No | |
| • | Conclusion, internal, construct, external | Internal, external | |
| Artifact | Availability | Yes | |
| | | | |

No

Paper Software visualization and deep transfer learning for effective software defect prediction

ID AP13

Experiments 5 Although the first one could be 3

| Aspect | Element | E1 | Comments |
|---------------------|---|-----------------------------------|--|
| Experiment type | | Optimization | Different hyperparameters |
| Hypotheses | Research hypotheses | No | No RQ |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 10 | Uses as base AlexNet network structure |
| | | #neurons/layer: No | Coneections described for non-AlexNet layers |
| | | connections: Partially | Params initialization is given only for AlexNet (as it |
| | | activation functions: Partially | is pre-trained) |
| | | params. Initialization: Partially | Some activation functions only (perhaps not all |
| | | paramor micanzación i arcany | layers have activation function) |
| | Model parameters | biases: No | Not mentioned in the paper |
| | Woder parameters | | Not mentioned in the paper |
| | 51 J 11 | weights: No | 5 |
| | DL algorithm | representation: Yes | Data augmentation is used for regularization |
| | | model type: Yes | |
| | | loss function: Yes | |
| | | regularization: Yes | |
| | | optimization: No | |
| | Training hyperparameters | train-test split:Yes | AlexNet is pre-trained with ImageNet 2012 |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 500 | |
| | Training data | yes | Explicitly linked |
| Operationalization | Factors and treatments | hyperparms (levels given) | Not sure if could be 3 experiments (with 2 factors) |
| | | 71-1-1 | or 2 experiments (with 3 factors) or 6 experiments |
| | | | with one factor |
| | Response variable, elaboration and metric | Yes | Fmeasure |
| Design | Design type | No | |
| _ | Blocking variables | Yes | Version perhaps? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | 10 | |
| Instrumentation | Test set | Yes | Explicitly linked |
| | Measuring instruments | No | • |
| | Measurement procedure | No | |
| | Technological infrastructure | Partial | Modern-day Linux server with 3 Titan XP GPUs |
| Population | Objects (chars. of the experimental datasets) | Yes | , |
| | | Partially | Average of 10 runs is reported only |
| | Descriptive statistics | | |
| Analysis | Descriptive statistics Inferential statistics | • | , , |
| Analysis | Inferential statistics | No | |
| | Inferential statistics | • | Internal, external, construct |

Paper Software visualization and deep trans

ID AP13 Experiments 5

| Aspect | Element | E2 | Comments |
|---------------------|---|--------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | |
| | | | |
| | Model parameters | Same as previous? | |
| | DL algorithm | Same as previous? | |
| | Training hyperparameters | Same as previous? | Train-test split: two consecutive versions of the |
| | | · | same project (this is done 2 times) |
| | Training data | Yes | Explicitly linked |
| Operationalization | Factors and treatments | NN model: Semantic, LSTM, CNN, | The other models are never explained. Mentions |
| | | DTL-DP | that same as reported in original papers |
| | Response variable, elaboration and metric | Yes | Fmeasure |
| Design | Design type | No | |
| | Blocking variables | Yes | Version perhaps? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | 10 | |
| Instrumentation | Test set | Yes | Explicitly linked |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partial | Modern-day Linux server with 3 Titan XP GPUs |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average of 10 runs is reported only |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Internal, external, construct |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Software visualization and deep trans

ID AP13 Experiments 5

| Aspect | Element | E3 | Comments |
|---------------------|---|----------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous? | |
| | model parameters | came as previous. | |
| | DL algorithm | Same as previous? | |
| | | | |
| | | | |
| | | | |
| | | | - · · · · · · · · · · · · · · · · · · · |
| | Training hyperparameters | Same as previous? | Train-test split: two different projects |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Explicitly linked |
| Operationalization | Factors and treatments | NN model: DBN, LSTM, CNN, DTL-DP | The other models are never explained. |
| | | | Mentions that same as reported in |
| | | | original naners |
| | Response variable, elaboration and metric | Yes | Fmeasure |
| Design | Design type | No | |
| | Blocking variables | Yes | Version perhaps? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | 22 | Not sure if I interpreted correctly |
| Instrumentation | Test set | Yes | Explicitly linked |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partial | Modern-day Linux server with 3 Titan XP GPU |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average of 22 runs is reported only |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Internal, external, construct |
| | A 11 1 1111 | V | |
| Artifact | Availability | Yes | |

Paper Software visualization and deep trans

ID AP13 Experiments 5

| Comments | , |
|----------|---|
| | |

| Aspect | Element | E4 | Comments |
|---------------------|---|---|---|
| Experiment type | | Optimization | incorporating layers |
| Hypotheses | Research hypotheses | Partially | RQ3 is really 2 experiments (E4 and E5) |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | |
| | | | |
| | Model parameters | Same as previous? | |
| | DL algorithm | Same as previous? | |
| | Training hyperparameters | Same as previous? | Train-test split: two consecutive versions of the same project (this is done 2 times) |
| Operationalization | Training data Factors and treatments | Yes DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) | Explicitly linked |
| | Response variable, elaboration and metric | Yes | Fmeasure, time |
| Design | Design type | No | |
| | Blocking variables | No | Version perhaps? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | 6? | Not clear. Cannot be deduced |
| Instrumentation | Test set | Yes | Explicitly linked |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partial | Modern-day Linux server with 3 Titan XP GPUs |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average of 10 runs is reported only |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Internal, external, construct |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Software visualization and deep trans

ID AP13 Experiments 5

Comments

| Aspect | Element | E5 | Comments |
|--|---|---|---|
| Experiment type | | Optimization | Incorporating layers |
| Hypotheses | Research hypotheses | Partially | RQ3 is really 2 experiments (E4 and E5) |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | |
| | | | |
| | Model parameters | Same as previous? | |
| | DL algorithm | Same as previous? | |
| | | | |
| | Training hyperparameters | Same as previous? | Train-test split: two different projects |
| | Training data | Yes | Explicitly linked |
| Operationalization | | | Explicitly linked |
| Operationalization | Factors and treatments | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) | Expiretty mixed |
| Operationalization | | DTL-DP architecture (Base, +TL, | Fmeasure, time |
| | Factors and treatments | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) | |
| | Factors and treatments Response variable, elaboration and metric | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes | |
| | Factors and treatments Response variable, elaboration and metric Design type | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No | |
| | Response variable, elaboration and metric Design type Blocking variables | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No | |
| Operationalization Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No | Fmeasure, time |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No No O O O O O O O O O O O O O | Fmeasure, time Not clear. Cannot be deduced |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No No No Yes | Fmeasure, time Not clear. Cannot be deduced |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No | Fmeasure, time Not clear. Cannot be deduced |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No | Not clear. Cannot be deduced Explicitly linked |
| Design Instrumentation | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No Partial | Not clear. Cannot be deduced Explicitly linked |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No No No Partial Yes | Not clear. Cannot be deduced Explicitly linked Modern-day Linux server with 3 Titan XP GPU |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No No Partial Yes Partially | Not clear. Cannot be deduced Explicitly linked Modern-day Linux server with 3 Titan XP GPU |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP) Yes No No No No No No No No Partial Yes Partially No | Not clear. Cannot be deduced Explicitly linked Modern-day Linux server with 3 Titan XP GPU Average of 10 runs is reported only |

Paper DLFix: Contexgt-based code transformation learning for automated program repair

ID AP14 Experiments 3

Comments I do not count their first experiment (compares against no-DNN). I divide in two their second experiment, as it in

| Aspect | Element | E1 | Comments |
|-------------------------------|---|------------------------------------|--|
| Experiment type | | Optimization | Embedded in E2 in paper. Hidden experiment |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | |
| | | #neurons/layer: No | |
| | | connections: No | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | biases: | biases: No | |
| | weights: | weights: No | |
| | DL algorithm | representation: Yes | |
| | 5 0 1 | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Mentions "epoch size" I understand this is #epochs |
| | Training Try perparameters | learning rate: Yes | (not batch size) |
| | | #iterations: No | (Not batter size) |
| | | batch size: No | |
| | | | |
| | Turbitus data | #epochs: Yes | Data Palanda antifant |
| 0 | Training data | Yes | Data linked to artifact |
| Operationalization | Factors and treatments | word2vec vector length (100, 50, | Decree as well word |
| | | 120), learning rate (0.001, 0.005, | Beam search used |
| | | 0.01), epoch size (100, 200, 300) | |
| | Response variable, elaboration and metric | Yes | Top K (k=1,5,10) |
| Design | Design type | No | 5 |
| | Blocking variables | Yes | Dataset could be (or may be factor) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | E haven |
| | Task duration | Yes No | 5 hours Looks like 1 run |
| Instrumentation | Number of experimental units Test set | Yes | Linked |
| mstramentation | | No | Linked |
| | Measuring instruments Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | They just reference the other two. Do not explain |
| Opulation | Objects (chars. of the experimental datasets) | 140 | |
| Analysis | Descriptive statistics | No | well their new one |
| Analysis | • | No | |
| Validity evaluation | Inferential statistics Conclusion internal construct external | Partially | Present a list of threats. No division |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Yes | riesent a list of threats. NO division |
| Aitilact | Badge | No | |
| | buuge | 140 | |

Paper DLFix: Contexgt-based code transform

ID AP14 Experiments 3

Comments I do not count their first experiment (cludes hyperparameter tuning

| Aspect | Element | E2 | Comments |
|---------------------|---|--------------------------------|---|
| Experiment type | | Evaluation | Mixed with E1 |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | biases: | Same as previous | |
| | weights: | | |
| | DL algorithm | Same as previous? | |
| | | | |
| | Training hyperparameters | Same as previous | Learning rate, wrod2vec vector lenght and epoch size chosen are not described |
| Operationalization | Training data Factors and treatments | Yes Neural model (Ratchet, | Data linked to artifact Do not mention anything about them. Jus that |
| ., | | Tufano(18), CODIT, Tufano(19)) | they had to implement CODIT, as code was not available |
| | Response variable, elaboration and metric | Yes | Top K (k=1,5,10) |
| Design | Design type | No | , , , , |
| | Blocking variables | No | Dataset could be (or may be factor) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | Yes | 5 hours |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Yes | Linked |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | They just reference the other two. Do not explain |
| | • | | well their new one |
| Analysis | Descriptive statistics | No | |
| , | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Present a list of threats. No division |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper DLFix: Contexgt-based code transform

ID AP14 Experiments 3

Comments I do not count their first experiment (c

| Aspect | Element | E3 | Comments |
|---------------------|---|---|---|
| Experiment type | | Optimization | incorporating layers/parts |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | biases: weights: | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | Learning rate, wrod2vec vector lenght and epoch size chosen are not described. Not sure if train-test split is the same |
| | Training data | Yes | Data linked to artifact |
| Operationalization | Factors and treatments | layer (Seq2Seq, Seq2Seq+PAT, 2layer- EDM, 2layer-EDM+PAT, 2layer- EDM+PAT+Re-ranking) | |
| | Response variable, elaboration and metric | Yes | Top1 |
| Design | Design type | No | |
| | Blocking variables | Yes | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Linked |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | They just reference the other two. Do not explain well their new one |
| Analysis | Descriptive statistics | No | |
| , | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Present a list of threats. No division |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | <u> </u> | | |

Paper Detection of hidden feature requests from massive chat messages via deep siamese network

ID AP15 Experiments 3

Comments Not sure the extent of the proposed solution. Mention data preparation, and preprocessing is one step

| Aspect | Element | E1 | Comments |
|---------------------|---|--|--|
| Experiment type | | Comparison with SOTA and optimization (n | · |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| Variables as last' | Statistical hypotheses | No #lover: 214/v2\1111 | 2 first layors are COTA discrete releases |
| variables selection | Model hyperparameters | #layers: 2+4(x2)+1+1 | 2 first layers are SOTA disentanglement |
| | | #neurons/layer: l1, l2 (512) | (already trained). Not clear if this is approach |
| | | connections: I1, I2 (feedforward), I3 (input), | |
| | | I4 (convolutional), I5 (BiLSTM), I6 | All descriptions are partial. |
| | | (combination), I7 (similarity) | Grid search used for: POS tag embedding |
| | | activation functions: I1, I2 softsign | (50), kernel sizes (2,3,4,5) feature |
| | | params. Initialization: l1, l2 (trained) | maps/kernel (25), output dimension of |
| | | | BiLSTM is 300 (150 for each direction). I |
| | | | assume these are the "chosen" but do not |
| | | | know the initial ones |
| | Model parameters | biases: No | Although I1 and I2 corresponds to a SOTA |
| | • | weights: No | disentanglement NN, and they are available |
| | DL algorithm | representation: Partially | It is not clear if their approach includes |
| | 22480 | model type: Partially | disentanglement or not. |
| | | loss function: Cross-entropy | They mention dropout and early stopping as |
| | | regularization: Dropout (0.1) and early | optimization, but it seems to me they are |
| | | | • |
| | | stopping (after 10 epochs) | regularization |
| | Training hunares | optimization: No | 2 fold intro project areas well-device from 2 |
| | Training hyperparameters | train-test split: Yes | 3-fold intra-project-cross-validation from 3 |
| | | learning rate: No | projects |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Explicitly linked to artifact |
| Operationalization | Factors and treatments | DNN model | FRMiner,, p-FRMiner, CNC, FT |
| | | | Others not explained (only p-FR miner). For |
| | | | CNC, codes and models provided in the |
| | | | publication. For FT, official released |
| | | | packages, trained (100 epochs, initial learning |
| | | | rate 1.0, n-gram 2), and hyperparameters |
| | | | tuning . |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | Deduced | Project? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No No | |
| | Procedure Number of experimental units | No Yes | Cross-validation |
| Instrumentation | Test set | Yes | Same as training data |
| | Measuring instruments | No | - |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Allennlp (open-source NLP library built on |
| | | | PyTorch). Missing versions |
| | | | NVIDIA 1060 GPU, intel core i7, 16GB RAM, |
| | | | Ubuntu |
| Population | Objects (chars. of the experimental datasets) | No | A table with some info is given, but nothing is |
| Analysis | Descriptive statistics | Partially | Average reported from 3-fold-cross-val |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | External, internal, construct (the authors |
| | | | define it for RV only) |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |
| | | | |

Paper Detection of hidden feature requests 1

ID AP15 Experiments 3

Comments Not sure the extent of the proposed so

| Aspect | Element | E2 | Comments |
|---------------------|--|-------------------------|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | Same for FRminer and p-Frminer |
| | Model parameters | Same as previous? | Comp for Epolica and a Espirar |
| | DL algorithm | Same as previous? | Same for FRminer and p-Frminer |
| | Training hyperparameters | Same as previous | Same for FRminer and p-Frminer |
| | Training data | Same as previous? | Same for FRminer and p-Frminer |
| Operationalization | Factors and treatments | DNN model, dataset size | FRMiner, p-FRMiner |
| | | | (initial, x5, x10, x20, x30) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | rredsion, recall, r 1 |
| 3 | Blocking variables | Deduce | Project? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Yes | Cross-validation |
| Instrumentation | Test set | Yes | Same as training data |
| | Measurement procedure | No No | |
| | Measurement procedure Technological infrastructure | No Same as previous | |
| | reamonogical iliitasuuctule | Same as previous | |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| Analysis | Descriptive statistics Inferential statistics | Partially No | Average reported from 3-fold-cross-val |
| Validity evaluation | Conclusion, internal, construct, external | Partially | External, internal, construct (the authors define it for RV only) |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |

Paper Detection of hidden feature requests f

ID AP15 Experiments 3

Comments Not sure the extent of the proposed so

Badge

| Aspect | Element | E3 | Comments |
|---------------------|---|---|---|
| Experiment type | | Comparison with SOTA and optimizati | ion (non-siamse version) |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous? | |
| | Model parameters | Same as previous? | |
| | DL algorithm | Same as previous? | |
| | Training hyperparameters | train-test split: Yes learning rate: No #iterations: No batch size: No | 3-fold cross-project-cross-validation from 3 projects |
| | | #epochs: No | |
| | Training data | Same as previous? | |
| Operationalization | Factors and treatments | DNN model | Same as E1 |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | , |
| | Blocking variables | Deduced | Project? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Yes | Cross-validation |
| Instrumentation | Test set | Yes | Same as training data |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| | | 2 | |
| Analysis | Descriptive statistics | Partially | Average reported from 3-fold-cross-val |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | External, internal, construct (the authors define it for RV only) |
| Artifact | Availability | Yes | |
| | Radge | Yes | Available |

Yes

Available

Paper RetrievaL-based neural source cocde summarization

ID AP16

Experiments 5 E5 is difficult to assess with this template. E4 encompasses 2 RQs

omments Some "hidden" experiments. Experiments are designed "on the fly" (during results & discussion).

| Aspect | Element | E1 | Comments |
|---|--|---------------------------------|--|
| Experiment type | | Optimization | "Hidden" experiment. Embedded in E2 |
| Hypotheses | Research hypotheses | No | |
| | | | |
| Variables selection | Statistical hypotheses | No Hayara Vas | hidden states in LCTM—several |
| variables selection | Model hyperparameters | #layers: Yes | hidden states in LSTM=several |
| | | #neurons/layer: No | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | No menciona regularization. |
| | | model type: Yes | Lambda = several |
| | | loss function: Yes | beam size = several |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | | |
| | - | #epochs: No | |
| 0 | Training data | Yes | Factors and the deduced to the contract of |
| Operationalization | Factors and treatments | | Factors can be deduced, treatments are not |
| | | hidden states=512, batch | explained "selecting the best one among |
| | | size=32, maximum | some alternatives" |
| | | iterations=100k, Adam | |
| | | optimizer, learning rate=0.001, | |
| | | beam size=5, lambda=3 | |
| | | | |
| | Response variable, elaboration and metric | Yes | BLEU(1-4), METEOR, ROUGE-L, CIDER |
| | | | |
| Design | Design type | No | |
| | Blocking variables | No | Dataset could be (or factor) |
| | Held-constant variables | No | , , |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | Mentions another paper |
| | Measurement procedure | No | Mentions another paper |
| | incusurement procedure | | Wendons another paper |
| | Technological infrastructure | Yes | Implemented using OpenNMT. |
| | | | Ubuntu 16.04 server with 16 cores of |
| | | | |
| | | | 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with |
| | Objects (shows of the constitution of the | Vec | 12GB memory |
| | Objects (chars. of the experimental datasets) | Yes | |
| | Descriptive statistics | No | |
| | Descriptive statistics | | |
| Analysis | Inferential statistics | No | Procent a list of throats. No division |
| Analysis Validity evaluation | Inferential statistics Conclusion, internal, construct, external | No Partially | Present a list of threats. No division |
| Population Analysis Validity evaluation Artifact | Inferential statistics | No | Present a list of threats. No division |

ID AP16 Experiments 5

Comments Some "hidden" experiments. Experiments

| Aspect | Element | E2 | Comments |
|---|---|------------------------------|---|
| Experiment type | | Evaluation | E1 is embedded here |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | | | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | Embedding size=256 |
| | | #neurons/layer: Yes | hidden states in LSTM=512 |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | No menciona regularization. |
| | | model type: Yes | Lambda = 3 |
| | | loss function: Yes | beam size = 5 |
| | | regularization: No | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: Yes | |
| | | learning rate: 0.001 | |
| | | #iterations: 100k | |
| | | batch size: 32 | |
| | | #epochs: No | |
| | Training data | Yes | |
| Operationalization | Factors and treatments | Neural model (CODE-NN, TL- | Does not mention anything about the others. |
| | | Codesum, Hybrid-DRL, GRNMET, | Proposal is partially explained |
| | | Rencos) | riopesaris paradily explained |
| | Response variable, elaboration and metric | Yes | BLEU(1-4), METEOR, ROUGE-L, CIDER |
| Design | Design type | No | |
| JC3IGI1 | Blocking variables | No | Dataset could be (or factor) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Number of experimental units | No | Looks like 1 run |
| nstrumentation | Test set | Yes | |
| | | | |
| | Measuring instruments | No | Mentions another paper |
| | Measurement procedure | No | Mentions another paper |
| | | | |
| | Technological infrastructure | Yes | Implemented using OpenNMT |
| | Technological infrastructure | Yes | Implemented using OpenNMT. |
| | Technological infrastructure | Yes | Ubuntu 16.04 server with 16 cores of 2.4GHz |
| | Technological infrastructure | Yes | Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Population | Technological infrastructure Objects (chars. of the experimental datasets) | | Ubuntu 16.04 server with 16 cores of 2.4GHz |
| • | | Yes Yes No | Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| • | Objects (chars. of the experimental datasets) | Yes | Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Analysis | Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Yes No | Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Population Analysis Validity evaluation Artifact | Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Yes No No | Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory |

ID AP16 Experiments 5

Comments Some "hidden" experiments. Experiments.

| Aspect | Element | E3 | Comments |
|---------------------------------------|---|---|--|
| Experiment type | | Optimization | Components |
| Hypotheses | Research hypotheses | Yes | RQ2, RQ4. Although they show them as |
| | | | two experiments |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DI algorithms | Comp or proving | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | |
| | | Component (NMT, NMT+syntactic, | |
| | | NMT+semantic, NMT+both) | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Yes | BLEU(1-4), METEOR, ROUGE-L, CIDER, |
| | | | For NMT and NM+both(Rencos), number |
| | | | of low-frequency words correctly |
| | | | generated |
| Design | Design type | No | |
| | Blocking variables | No | Dataset could be (or factor) |
| | Held-constant variables | No | |
| | | No | |
| | Measured variables (covariates) | | |
| | Randomization | No | |
| | Randomization Task duration | No No | Looks like 1 run |
| Instrumentation | Randomization Task duration Number of experimental units | No No | Looks like 1 run |
| Instrumentation | Randomization Task duration | No No | Looks like 1 run |
| Instrumentation | Randomization Task duration Number of experimental units Test set | No No No Yes | Looks like 1 run |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments | No No No Yes | |
| Instrumentation | Randomization Task duration Number of experimental units Test set | No No No Yes | For number of low-frequency words |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments | No No No Yes | For number of low-frequency words correctly generated it is partially |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No Yes No Partially | For number of low-frequency words correctly generated it is partially explained |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments | No No No Yes | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No Yes No Partially | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of |
| Instrumentation | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No Yes No Partially | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU |
| | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No Yes No Partially | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of |
| Population | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No Yes No Partially Yes | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU |
| Instrumentation Population Analysis | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No Yes No Partially Yes Yes No | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU |
| Population | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No Yes No Partially Yes | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU |
| Population Analysis | Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No Yes No Partially Yes No No No | For number of low-frequency words correctly generated it is partially explained Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory |

ID AP16 Experiments 5

Comments Some "hidden" experiments. Experime

| Aspect | Element | E4 | Comments |
|---------------------------------------|---|---|---|
| Experiment type | | Optimization | Components not part of DNN |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | | | |
| Variables selection | Statistical hypotheses | No Sama as provious | |
| variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | <u> </u> | • | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | |
| | | k (110), T (0, 0.2, 0.5, 0.8) | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Yes | AVERAGED BLEU, METEOR, ROUGE-L, CIDER |
| | Response variable, elaboration and metric | Yes | AVERAGED BLEU, METEOR, ROUGE-L, CIDER |
| | Response variable, elaboration and metric | Yes | AVERAGED BLEU, METEOR, ROUGE-L, CIDER |
| Design | | | AVERAGED BLEU, METEOR, ROUGE-L, CIDER |
| Design | Design type | No | |
| Design | | | AVERAGED BLEU, METEOR, ROUGE-L, CIDER Dataset could be (or factor) |
| Design | Design type Blocking variables | No No | |
| Design | Design type Blocking variables Held-constant variables | No No No | |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No No No No | |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units | No No No No No No | |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No No No No No | Dataset could be (or factor) |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set | No No No No No No No Vo | Dataset could be (or factor) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | No No No No No No No No | Dataset could be (or factor) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set | No No No No No No No Vo | Dataset could be (or factor) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | No No No No No No No No | Dataset could be (or factor) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No No No Yes | Dataset could be (or factor) Looks like 1 run |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments | No No No No No No No No | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No No No Yes | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No No No Yes | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No No No Yes | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No No No No No No Yes Yes | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No No No No No Yes Yes No | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No No No No No Yes Yes No No | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No No No No No Yes Yes No | Dataset could be (or factor) Looks like 1 run Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB |

ID AP16 Experiments 5

Comments Some "hidden" experiments. Experiments.

| Aspect | Element | E5 | Comments |
|--------------------|---|---|--|
| Experiment type | - 102 | Evaluation | Humans assess similarity between proposed and re |
| lypotheses | Research hypotheses | No | |
| | | | |
| | Statistical hypotheses | No | |
| ariables selection | Model hyperparameters | Same as previous? | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous? | |
| | | | |
| | DL algorithm | Same as previous? | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous? | |
| | | | |
| | | | |
| | | | |
| | Training data | Como os provinus | |
| nerationalization | Training data Factors and treatments | Same as previous Neural model (Hybrid-DRL, Rencos) | |
| perationalization | ractors and deadnerits | Neural model (Hybrid-DKL, Kericos) | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Yes | Similarity as Likert scale (1-5) |
| | nesponse vanasie, clasoration and metre | 163 | Similarity as Electrocate (1 5) |
| | | | |
| | | | |
| esign | Design type | No | |
| - | Blocking variables | No | Dataset could be (or factor) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | Yes | subjects assigned to samples |
| | Task duration | No | |
| | Number of experimental units | Yes | Each sample ranked by 3 people |
| strumentation | Test set | Partially | Do not mention which samples chosen. |
| | | | People from Amazon MT |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | | | |
| | | | |
| | Technological infrastructure | Yes | |
| | | | |
| | | | |
| | | | |
| opulation | Objects (chars. of the experimental datasets) | Yes | |
| nalysis | Descriptive statistics | Yes | |
| | Inferential statistics | Partially | Not sure if correct. Wilcoxon |
| | Conclusion, internal, construct, external | Partially | Present a list of threats. No division |
| rtifact | Availability | Yes | |
| | Badge | No | |

Paper On learning meaningul assert statements for unit test cases

ID AP17 Experiments 3

Comments E1 is hidden

| Speciment type Optimization Hypotheses Yes RQ1 | Aspect | Element | E1 | Comments |
|--|---------------------|---|--------------|----------------------------------|
| Statistical hypotheses | Experiment type | | Optimization | |
| Variables selection Model hyperparameters Same as next Model parameters Same as next DL algorithm Same as next Training hyperparameters Same as next Training hyperparameters Same as next Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Packet Same as next Design Vpe No No Held-constant variables No No Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Task duration No No Procedure No No Number of experimental units No | Hypotheses | Research hypotheses | Yes | RQ1 |
| Model parameters DL algorithm Same as next DL algorithm Same as next Training hyperparameters Same as next Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design blocking variables Response variables No Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Task duration No Procedure No Nomber of experimental units No Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars, of the experimental datasets) Portally Inclusion/exclusion Analysis Descriptor statistics No Validity evaluation Outsion, internal, construct, external Partially missing conclusion | | Statistical hypotheses | No | |
| Design Design type Response variable, elaboration and metric Yes Raw source and abstracted code Design Design type Response variable, elaboration and metric Yes Loss function Design Design type No Held-constant variables No Held-constant variables No Measured variables No Randomization No Randomization No Procedure No No No Head No Procedure No | Variables selection | Model hyperparameters | Same as next | |
| Design Design type Response variable, elaboration and metric Yes Raw source and abstracted code Design Design type Response variable, elaboration and metric Yes Loss function Design Design type No Hed-constant variables No Hed-constant variables No Measured variables No Randomization No Procedure No Procedure No No No Head No Procedure No | | | | |
| Design Design type Response variable, elaboration and metric Yes Raw source and abstracted code Design Design type Response variable, elaboration and metric Yes Loss function Design Design type No Hed-constant variables No Hed-constant variables No Measured variables No Randomization No Procedure No Procedure No No No Head No Procedure No | | | | |
| Design Design type Response variable, elaboration and metric Yes Raw source and abstracted code Design Design type Response variable, elaboration and metric Yes Loss function Design Design type No Held-constant variables No Held-constant variables No Measured variables No Randomization No Randomization No Procedure No No No Head No Procedure No | | | | |
| Training hyperparameters Same as next Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design type Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization No Task duration Procedure No No Nomeroredure No Nomeroredure No Nomeroredure No Measurement procedure No Measurement procedure No Measurement procedure No Test set Measurement procedure No Measurement procedure No Descriptive statistics No Population Objects (chars. of the experimental datasets) Partially Analysis Descriptive statistics No Validity evaluation Conclusion, internal, construct, external Partially Missing conclusion Artifact Available No missing conclusion | | Model parameters | Same as next | |
| Training data Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Pesign Design type Response variables, elaboration and metric No Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization Task duration No Procedure No Number of experimental units No Instrumentation Test set Measuring instruments No Measurement procedure No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Validity evaluation Conclusion, internal, construct, external Partially Meissing conclusion Artifact Available in RP Reswource and abstracted code Available in RP No Held-constant variables No Included in RP Raw source and abstracted code Available in RP No Held-constant variables No Included in RP Inclusion/exclusion Inclusion/exclusion Partially Missing conclusion Artifact Availablity Yes | | DL algorithm | Same as next | |
| Training data Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Pesign Design type Response variables, elaboration and metric No Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization Task duration No Procedure No Number of experimental units No Instrumentation Test set Measuring instruments No Measurement procedure No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Validity evaluation Conclusion, internal, construct, external Partially Meissing conclusion Artifact Available in RP Reswource and abstracted code Available in RP No Held-constant variables No Included in RP Raw source and abstracted code Available in RP No Held-constant variables No Included in RP Inclusion/exclusion Inclusion/exclusion Partially Missing conclusion Artifact Availablity Yes | | | | |
| Training data Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Pesign Design type Response variables, elaboration and metric No Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization Task duration No Procedure No Number of experimental units No Instrumentation Test set Measuring instruments No Measurement procedure No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Validity evaluation Conclusion, internal, construct, external Partially Meissing conclusion Artifact Available in RP Reswource and abstracted code Available in RP No Held-constant variables No Included in RP Raw source and abstracted code Available in RP No Held-constant variables No Included in RP Inclusion/exclusion Inclusion/exclusion Partially Missing conclusion Artifact Availablity Yes | | | | |
| Training data Training data Yes Raw source and abstracted code Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Pesign Design type Response variables, elaboration and metric No Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization Task duration No Procedure No Number of experimental units No Instrumentation Test set Measuring instruments No Measurement procedure No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Validity evaluation Conclusion, internal, construct, external Partially Meissing conclusion Artifact Available in RP Reswource and abstracted code Available in RP No Held-constant variables No Included in RP Raw source and abstracted code Available in RP No Held-constant variables No Included in RP Inclusion/exclusion Inclusion/exclusion Partially Missing conclusion Artifact Availablity Yes | | | | |
| Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design type No Blocking variables No Held-constant variables Held-constant variables (covariates) No Held-constant variables (covariates) Randomization No Held-constant variables (covariates) Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Held-constant variables | | Training hyperparameters | Same as next | |
| Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design type No Blocking variables No Held-constant variables Held-constant variables (covariates) No Held-constant variables (covariates) Randomization No Held-constant variables (covariates) Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Held-constant variables | | | | |
| Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design type No Blocking variables No Held-constant variables Held-constant variables (covariates) No Held-constant variables (covariates) Randomization No Held-constant variables (covariates) Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Held-constant variables | | | | |
| Operationalization Factors and treatments Yes Available in RP Response variable, elaboration and metric Yes Loss function Design Design type No Blocking variables No Held-constant variables Held-constant variables (covariates) No Held-constant variables (covariates) Randomization No Held-constant variables (covariates) Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Hour Task duration No Held-constant variables (covariates) No Held-constant variables (covariates) No Held-constant variables | | Tutting day | Was | Per conservate leaders desidente |
| Design Design type No Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization No Procedure No No Number of experimental units No Measurement procedure No | 0 | | | |
| Design Design type No No Held-constant variables No Measured variables No Measured variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Measured variables (covariates) No Measured variables (covariates) No Mo House of experimental units No Measured procedure No Measured procedure No Measurement Procedur | Operationalization | Factors and treatments | Yes | Available in RP |
| Design Design type No No Held-constant variables No Measured variables No Measured variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Measured variables (covariates) No Measured variables (covariates) No Mo House of experimental units No Measured procedure No Measured procedure No Measurement Procedur | | | | |
| Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Ves Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Availability Yes | | Response variable, elaboration and metric | Yes | Loss function |
| Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Ves Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Availability Yes | | | | |
| Blocking variables No Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Ves Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Availability Yes | | | | |
| Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No | Design | Design type | No | |
| Measured variables (covariates) No Randomization No Task duration No Procedure No No No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Availability Yes | | Blocking variables | No | |
| Randomization No Task duration No Procedure No Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Held-constant variables | No | |
| Task duration No Procedure No Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Measured variables (covariates) | No | |
| Procedure No No Number of experimental units No No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Randomization | No | |
| Number of experimental units No Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Task duration | No | |
| Instrumentation Test set Yes included in RP Measuring instruments No Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Procedure | No | |
| Measuring instruments Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion | | Number of experimental units | No | |
| Measurement procedure No Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | Instrumentation | Test set | Yes | included in RP |
| Technological infrastructure No Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Measuring instruments | No | |
| Population Objects (chars. of the experimental datasets) Partially Inclusion/exclusion Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Measurement procedure | No | |
| Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | | Technological infrastructure | No | |
| Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | Population | Objects (chars. of the experimental datasets) | Partially | Inclusion/exclusion |
| Validity evaluation Conclusion, internal, construct, external Partially missing conclusion Artifact Availability Yes | Analysis | | | |
| Artifact Availability Yes | | | | |
| · | | | | missing conclusion |
| Badge No | Artifact | | | |
| | | Badge | No | |

Paper On learning meaningul assert stateme

ID AP17 Experiments 3

Comments E1 is hidden

| Aspect | Element | E2 | Comments |
|---------------------|---|-----------------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ3, R4, RQ6 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 6? | |
| | | #neurons/layer: No | |
| | | connections: input, bidirectional | |
| | | RNN (2xLSTM), attention, 2xLSTM | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Model is encoder/attention/decoder |
| | | model type: Yes | |
| | | loss function: negative log | |
| | | likelihood | |
| | | regularization: encoder & decoder | |
| | | dropout (0.2) | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | Training hyperparameters | learning rate: 0.0001 | |
| | | • | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | In replication package |
| Operationalization | Factors and treatments | Partially | Assessment. No comparison against others. Compares raw/abstract models |
| | Response variable, elaboration and metric | Partially | # perfect predictions, BLEU-4 (for imperfect). |
| | | | BLEU not properly explained, complementarines |
| | | | raw/abstract models, # and % asserts resolved |
| | | | with copy mechanism, time to generated assert |
| | | | statements |
| Design | Design type | No | |
| · · | Blocking variables | Yes | Beam size (150 step 5) |
| | Held-constant variables | No | , , |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Included in RP |
| | Measuring instruments | No | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Partially | Inclusion/exclusion |
| Analysis | Descriptive statistics | No | |
| • | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | missing conclusion |
| Artifact | Availability | Yes | - |
| | Badge | No | |

Paper On learning meaningul assert stateme

ID AP17 Experiments 3

Comments E1 is hidden

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | Woder parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | In replication package |
| Operationalization | Factors and treatments | Partially | compares abstract model vs. Frequency model |
| | | | obtained from examination of most common predicted |
| | Response variable, elaboration and metric | Yes | predicted |
| | | | |
| | | | |
| | | | |
| Design | Design type | No | Page 252 /1 5 40) |
| | Blocking variables | Yes | Beam size (1, 5, 10) |
| | Held-constant variables Measured variables (covariates) | No No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | | |
| | Number of experimental units | No No | Seems 1 run |
| Instrumentation | Test set | Yes | Included in RP |
| msuumentauon | Measuring instruments | No | included in NF |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Partially | Inclusion/exclusion |
| Analysis | Descriptive statistics | No | mousiony exclusion |
| Alialy 313 | Inferential statistics | No | |
| Validity evaluation | | Partially | missing conclusion |
| Artifact | Availability | Yes | missing conclusion |
| | Badge | No | |
| | 0- | | |

Paper Automatic extraction of opinion-based Q&A from online developer chats

ID AP18 Experiments 1

Comments The first experiment is outside the DNN

| Aspect | Element | E1 | Comments |
|---------------------|---|---------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | GloVe for word embeddings and implement |
| | | #neurons/layer: 3 convolution | TextCNN for sentence encoding |
| | | filters (size 2, 3, 4) with 50 | |
| | | feature maps per filter. Pool | |
| | | sizes of convolution are (2,1), | |
| | | (2,1), (3,1). BiLSTM with 200x2 | |
| | | units. | |
| | | connections: Yes | |
| | | activation functions: sigmoid | |
| | | for linear layer. Rest (?) | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | • | weights: No | |
| | DL algorithm | representation: Yes | |
| | . 0 | model type: Yes | |
| | | loss function: binary-cross- | |
| | | entropy | |
| | | regularization: dropout: 0.5 | |
| | | (Text CNN embeddings), 0.1 | |
| | | (LSTM), early stopping | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 80-20 | Grid search for hyper-parameter tuning |
| | Training hyperparameters | learning rate: No | Glid Scarciffor Hyper-parameter turning |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Linked |
| Operationalization | Factors and treatments | Partially | Model type. The others are not SOTA, but |
| Operationalization | ractors and deatherts | rardany | created by the authors |
| | Response variable, elaboration and metric | Yes | With formulas |
| Design | Design type | No | With formulas |
| Design | Blocking variables | Deduced | Programming community |
| | | No | Find the state of |
| | Held-constant variables | | |
| | Measured variables (covariates) Randomization | No No | |
| | Task duration | | |
| | Procedure | No | |
| | | No | Cooms 1 mm |
| Instrumentation | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked |
| | Measurement procedure | No | |
| | Measurement procedure | No Portially | Vorse 2 ECHalatel Come IF 000 DDD2 DAA4 |
| Population | Technological infrastructure Objects (charse of the experimental datasets) | Partially | Keras, 2.5GHz Intel Core i5, 8GB DDR3 RAM |
| | Objects (chars. of the experimental datasets) | Yes | Explained |
| Analysis | Descriptive statistics | Partially | Average |
| Validity avaluati | Inferential statistics | No Partially | Internal construct external Militaria |
| validity evaluation | Conclusion, internal, construct, external | Partially | Internal, construct, external. Missing |
| A.+:6+ | A the health. | V | conclusion |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Automatically matching bug reports with related app reviews

ID AP19 Experiments 1

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|---|
| Experiment type | | Evaluation | Does not compare against anything. Just |
| | | | assesses |
| Hypotheses | Research hypotheses | Yes | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | DISTILBERT creates 768-dimensional |
| | | #neurons/layer: No | embeddings. Nothing is mention on whether |
| | | connections: No | defaults are used for its implementation |
| | | activation functions: No | (layers, units, etc.) or something is changed |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: No | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | No | |
| Operationalization | Factors and treatments | Partially | Assessment. No comparison |
| | Response variable, elaboration and metric | Yes | MAP, Hit ratio, #relevant/irrelevant |
| | | | matches, |
| Design | Design type | No | |
| | Blocking variables | Deduced | 4 apps used, number of suggestions |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | RP is explicitly linked |
| | Measuring instruments | Yes | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Mean and boxplots |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Internal and external only |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper CURE: Code-aware neural machine translation for automatic program repair

ID AP20 Experiments 3

Comments E1 is hidden. They acknowledge to have done random search for hyperparameters tuning, but do not expla

| Aspect Experiment type | Element | E1 Optimization | Comments |
|---|---|---|--|
| Hypotheses | Research hypotheses | No | |
| .,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: GPT model: 8 transformer blocks + 6 attention heads(?) #neurons/layer: 384 input connections: Yes activation functions: No params. Initialization: No | Transformer block: (self-attention+ normalization+feedforward+normalization) |
| | Model parameters | biases: No weights: No | |
| | DL algorithm | representation: Yes model type: Yes GPT model: loss function: Yes regularization: Dropout optimization: Adam APR model: loss function: Yes regularization: No optimization: Adam | There are 2 components: PL model (GPT, embeddings) and NMT model (CoNuT, 2 encoders+1 decoder+attention+token generator) Uses code-aware beam search (outside DNN) |
| | Training hyperparameters | GPT model: train-test split: Yes learning rate: 0-2.5e4 at the first 2,000 training steps, then decreases with cosine #iiterations: No batch size: 12 #epochs: 5" APR model: train-test split: Yes learning rate: 6.25e-5 #iterations: No batch size: 12 #epochs: 1 | There are 2 trainings: GPT-PL model (for embeddings), GPT-PL+NMT model APR task Ensemble learning |
| | Training data | Partially | Uses CoCoNuT, but makes modifications (explained). No explicit link to artifact |
| Operationalization | Factors and treatments | Yes | #convolution dimension (128-512), kernel size (2-10), number of onvolutional layers (1-5), dropout (0-0.5) Random search |
| | Response variable, elaboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | No | Not mentioned |
| | Measuring instruments | No | |
| | Measurement procedure | Partially | Beam search |
| | Technological infrastructure | Partially | GPT implemented by Huggin Face, CoNuT and Fconv, implemented using fairseq. 56-core server with 1 NVIDIA TITAN V and 3 Xp GPUs |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partially | Named limitations. Flat list |
| Artifact | Availability | Yes | Github |
| | Badge | No | |
| | Badge | No | |

CURE: Code-aware neural machine tra AP20 Paper

ID Experiments 3

Comments E1 is hidden. They acknowledge to havin more

| Aspect | Element | E2 | Comments |
|------------------------|---|-------------------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | biases: No | |
| | DL algorithm | weights: No Same as previous | |
| | Dealgonami | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | GPT model: | |
| | | train-test split: Yes | |
| | | learning rate: 0-2.5e4 at the first | |
| | | 2,000 training steps, then | |
| | | decreases with cosine | |
| | | #iterations: No | |
| | | batch size: 12 | |
| | | #epochs: 5 | |
| | | APR model: | |
| | | train-test split: Yes | |
| | | learning rate: 6.25e-5 | |
| | | #iterations: No | |
| | | batch size: 12 | |
| | | #epochs: 1 | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | CURE compared with 25 APR techniques (only 8 |
| o pera dorianza dori | ractors and decaments | randany | appear in paper) |
| | | | |
| | Response variable, elaboration and metric | Yes | #corrected bug, #plausible bugs, compilable |
| | | | rates, time (CURE only) |
| Design | Design type | No | |
| | Blocking variables | Deduced | Test set? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Uses Defects4J and QuixBugs. Reference to |
| | | | where taken |
| | Measuring instruments | No Postially | Dog w gooveh |
| | Measurement procedure | Partially | Beam search |
| | Technological infrastructure | Partially | GPT implemented by Huggin Face, CoNuT and |
| | | | Fconv, implemented using fairseq. 56-core server with 1 NVIDIA TITAN V and 3 Xp GPUs |
| Danisla Ka | Obi- de /abana afabana a a a a a a a a a a a a a a a a a | N- | |
| Population Analysis | Objects (chars. of the experimental datasets) Descriptive statistics | No Partially | 1 single datapoint per tool/test set for numbers. |
| , | , | | For compilable rates, average is presented (N |
| | | | datapoints) |
| v P.Pr | Inferential statistics | No | All the state of the state of |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Named limitations. Flat list |
| Artifact | Availability | Yes | Github |

Paper CURE: Code-aware neural machine tra

ID AP20 Experiments 3

Comments E1 is hidden. They acknowledge to hav

| Aspect | Element | E3 | Comments |
|---------------------|------------------------|------------------|----------|
| Experiment type | | Optimization | Ablation |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Same as previous | |
|---------------------|---|------------------|--|
| Operationalization | Factors and treatments | Partially | CURE compared with 4 more options removing components |
| | Response variable, elaboration and metric | Yes | #corrected bug, #plausible bugs, compilable rates, length of candidate patches, #OOV tokents |
| Design | Design type | No | |
| | Blocking variables | Deduced | Test set? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Uses Defects4J and QuixBugs. Reference to where taken |
| | Measuring instruments | No | |
| | Measurement procedure | Partially | Beam search |
| | Technological infrastructure | Partially | GPT implemented by Huggin Face, CoNuT and Fconv, implemented using fairseq. 56-core server with 1 NVIDIA TITAN V and 3 Xp GPUs |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | 1 single datapoint per tool/test set for numbers. For compilable rates, average is presented (N datapoints) |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Named limitations. Flat list |
| Artifact | Availability | Yes | Github |
| | Badge | No | |

Paper Automated query reformulation for efficient search based on query logs from stack overflow

ID AP21 Experiments 1

Comments Hyperparameters tuning done via grid search (no experiments on that)

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | Generic description of the structure of the |
| | | #neurons/layer: Some | layers, but no exact number of |
| | | connections: Yes | encoders/decoders |
| | | activation functions: Some | |
| | | params. Initialization: No | ReLU for output layer of last encoder |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | BPE (byte pari enconding. Do not mention if |
| | | model type: Transformer | a DNN is used for this) |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 80-10-10 | Beam search=10 |
| | | learning rate: 0.001 | Length normalization parameter alpha=0.6 |
| | | #iterations: No | |
| | | batch size: 256 | Grid search for hp tuning |
| | | #epochs: 147 | |
| | Training data | Partially | Explain how it has been obtained, but no |
| | | | explicit link |
| Operationalization | Factors and treatments | Partially | SEQUER, GEC, GooglePS, seq2seq, |
| | | | seq2seq_attn, HREDqs. Seq2seq and HRED |
| | | | are trained like SEQUER |
| | Response variable, elaboration and metric | Yes | EM (@1,5,10), GLEU, M ² (@P,R,F1), MRR |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | tensor2tensor library. 4 NVIDIA V100 GPU |
| | | | (32GB memory) |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | 1 single value |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | No | |
| Validity evaluation | | | |
| Artifact | Availability | Yes | |

Paper Fault localization with code coverage representation learning

ID AP22 Experiments 4

| Aspect | Element | E1 | Comments |
|---------------------|---|-----------------------------|--|
| Experiment type | | Optimization | Fine-tuning hidden |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | Word2Vec is used |
| | | #neurons/layer: No | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: L2, dropout | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | They test one bug by training on all other bugs |
| | · | learning rate: factor | , , , , , |
| | | #iterations: No | |
| | | batch size: factor | |
| | | #epochs: factor | |
| | Training data | Yes | Public data (Defects4J) |
| Operationalization | Factors and treatments | Yes | Epoch size (100, 200, 300), batch size (64, 128, 256); |
| | | | learning rate (0.001, 0.003, 0.005, 0.010); vector |
| | | | length of word representation (150, 200, 250, 300), |
| | | | convolutional core size (3x3, 5x5, 7x7, 9x9, 11x11); |
| | | | #convolutional cores (3, 5, 7, 9, 11) |
| | | | Word2Vec is also fine-tuned (epoch number, loss |
| | | | functions, learning rate) |
| | Response variable, elaboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | Deduced | cross-validation on faults for each individual project |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Partially | Cross-validation |
| Instrumentation | Test set | Yes | Public data (Defects4J) |
| | Measuring instruments | No | , |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | | Partially | Flat list of 2 threats |
| Artifact | Availability | Yes | |
| arace | Badge | No | |
| | paabe | 140 | |

ID AP22 Experiments 4

| Aspect | Element | E2 | Comments |
|--|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | We do not know the levels of the factors chosen |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | Ochiai, Dstar (no DNNs) MUSE, Metallaxis (no DNNs) RBF Neural network, DeepFL (DNNs SOTA) |
| | | | |
| | Response variable, elaboration and metric | Yes | Top-1, top-3, top-5, P%,MFR, MAR |
| Design | Response variable, elaboration and metric Design type | Yes No | Top-1, top-3, top-5, P%,MFR, MAR |
| Design | • | | Top-1, top-3, top-5, P%,MFR, MAR cross-validation on faults for each individual project |
| Design | Design type | No | cross-validation on faults for each individual |
| Design | Design type Blocking variables | No Deduced | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables | No Deduced | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No Deduced No No | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | No Deduced No No | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No Deduced No No No No | cross-validation on faults for each individual |
| · | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No Deduced No No No No | cross-validation on faults for each individual project |
| · | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No Deduced No No No No No Partially | cross-validation on faults for each individual project Cross-validation |
| · | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No Deduced No No No No Partially Yes | cross-validation on faults for each individual project Cross-validation |
| - | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No Deduced No No No No Partially Yes No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No Deduced No No No No Partially Yes No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No Deduced No No No No Partially Yes No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No Deduced No No No No Partially Yes No No No No No Yes | cross-validation on faults for each individual project Cross-validation |
| Design Instrumentation Population Analysis Validity evaluation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Deduced No No No No Partially Yes No No No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Deduced No No No No Partially Yes No | cross-validation on faults for each individual project Cross-validation Public data (Defects4J) |

ID AP22 Experiments 4

| Aspect | Element | E3 | Comments |
|---|---|---|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | We do not know the levels of the factors chosen |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | MULTRIC, FLUCCS, TraPT (no DNNs) DeepFL (DNN) |
| | | | |
| | Response variable, elaboration and metric | Yes | Top-1, top-3, top-5, P%,MFR, MAR |
| Design | Response variable, elaboration and metric Design type | Yes | Top-1, top-3, top-5, P%,MFR, MAR |
| Design | Response variable, elaboration and metric Design type Blocking variables | Yes No Deduced | Top-1, top-3, top-5, P%,MFR, MAR cross-validation on faults for each individual project |
| Design | Design type | No | cross-validation on faults for each individual |
| Design | Design type Blocking variables | No Deduced | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables | No Deduced | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No Deduced No No | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | No Deduced No No No | cross-validation on faults for each individual |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No Deduced No No No No | cross-validation on faults for each individual project Cross-validation |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No Deduced No No No No Partially Yes | cross-validation on faults for each individual project |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No Deduced No No No No Partially Yes No | cross-validation on faults for each individual project Cross-validation |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No Deduced No No No No Partially Yes No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No Deduced No No No No Partially Yes No No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No Deduced No No No No Partially Yes No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No Deduced No No No No Partially Yes No No No | cross-validation on faults for each individual project Cross-validation |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Deduced No No No No No Partially Yes No | cross-validation on faults for each individual project Cross-validation Public data (Defects4J) |
| Instrumentation Population Analysis Validity evaluation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No Deduced No No No No No Partially Yes No No No No Partially Yes Partially | cross-validation on faults for each individual project Cross-validation |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Deduced No No No No No Partially Yes No | cross-validation on faults for each individual project Cross-validation Public data (Defects4J) |

ID AP22 Experiments 4

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | We do not know the levels of the factors chosen |
| | Training data | Same as previous | |
| Орегинопиндиноп | Factors and treatments | Partially | Does not compare against SOTA. Merely assesses |
| | Decrease unitable alchematics and matrix | Voc | Top 1 DOWNED MAD |
| 5 . | Response variable, elaboration and metric | Yes | Top-1, P%,MFR, MAR |
| Design | Design type Blocking variables | No Deduced | cross-validation on faults for each individual project |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Partially | Cross-validation |
| Instrumentation | Test set | Yes | Public data (Defects4J) |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 2 threats |
| Artifact | Availability | Yes | |
| | Badge | No | |

Measurement procedure Technological infrastructure

Descriptive statistics

Inferential statistics

Validity evaluation Conclusion, internal, construct, external

Availability

Badge

Objects (chars. of the experimental datasets)

Population

Analysis

Artifact

ID AP22 Experiments 4

Comments R3 and RQ4 are outside DNN

| Aspect | Element | E5 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ6 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | We do not know the levels of the factors chosen |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | | |
| | | | |
| | Response variable, elaboration and metric | Yes | Top-1, P%,MFR, MAR |
| Design | Design type | No | |
| | Blocking variables | Deduced | cross-validation on faults for each individual project |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Partially | Cross-validation |
| Instrumentation | Test set | Yes | Public data (Defects4J) and ManyBugs |
| | Measuring instruments | No | |
| | | | |

No

No

Yes

No

No

Yes

No

Partially

Flat list of 2 threats

Paper Code prediction by feeding trees to transformers

ID AP23 Experiments 4

Comments E4 is included in related work, at the end of the paper. No associated RQ

| Aspect | Element | E1 | Comments |
|---------------------|---|--------------------------------|---|
| Experiment type | | Evaluation+optimization | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Partially | Pytorch implementation of GPT-2. 6 |
| | | #neurons/layer: No | transformer blocks, six heads per block, |
| | | connections: Yes | n_ctx =1000, embedding_dim = 300 |
| | | activation functions: No | |
| | | params. Initialization: Random | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: cross-entropy | |
| | | regularization: No | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: No | "Other hyperparameters" borrowed from |
| | | learning rate: 1e-3 | [20] |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: Yes (for all models) | |
| | Training data | Partially | Refers to "data preparation pipeline", not to |
| | | | data itself. Besides, they use 2 datasets: a |
| | | | Facebook internal repository and the py150 |
| | | | dataset available from other publication |
| | | | (which they modify) |
| Operationalization | Factors and treatments | Partially | SeqRNN, SeqTrans, TravTrans (type+value) |
| | | | (2 proposals, the differences are |
| | | | inputs/outputs only) |
| | Response variable, elaboration and metric | Yes | Training time (min/epoch), inference time, |
| | nesponse variable, elaboration and metric | 163 | model size, MRR |
| Design | Design type | No | model size, with |
| Design | Blocking variables | Deduced | type of token? Dataset |
| | Held-constant variables | No | type of tokeni batabet |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training test |
| | Measuring instruments | No | as as |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | NVIDIA Tesla V100, 4 GPUs |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | Overall/mean |
| Allalysis | Inferential statistics | No | Overally mean |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | i ide iist |
| , a diace | Badge | No | |
| | Dauge | 110 | |

Paper Code prediction by feeding trees to tra

ID AP23 Experiments 4

Comments E4 is included in related work, at the ϵ

| Aspect | Element | E2 | Comments |
|-------------------------------------|---|--|---|
| Experiment type | | Evaluation+optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| | Trailing data | Same as previous | |
| | | | |
| Operationalization | Factors and treatments | Deep3, Code2Seq, PathTrans, | |
| | | TravTrans (2 proposals, the | |
| | | | |
| | | differences are inputs/outputs only) | |
| | Response variable, elaboration and metric | differences are inputs/outputs only) Yes | Training time (min/epoch), inference time, model size, MRR |
| Design | Response variable, elaboration and metric Design type | | |
| Design | <u> </u> | Yes | |
| Design | Design type | Yes | size, MRR |
| Design | Design type Blocking variables | Yes No Deduced | size, MRR |
| Design | Design type Blocking variables Held-constant variables | Yes No Deduced No | size, MRR |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | Yes No Deduced No No | size, MRR |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | Yes No Deduced No No No | size, MRR |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | Yes No Deduced No No No No | |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | Yes No Deduced No No No No No | size, MRR |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | Yes No Deduced No No No No No No | size, MRR type of token? Dataset |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | Yes No Deduced No No No No No Partially No No | size, MRR type of token? Dataset |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | Yes No Deduced No No No No Partially No | size, MRR type of token? Dataset |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Yes No Deduced No No No No No No No No Partially No Partially No | size, MRR type of token? Dataset Same as training test NVIDIA Tesla V100, 4 GPUs |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Yes No Deduced No No No No No Partially No Partially | size, MRR type of token? Dataset Same as training test |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Yes No Deduced No No No No No No No No Partially No Partially No | size, MRR type of token? Dataset Same as training test NVIDIA Tesla V100, 4 GPUs |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics Conclusion, internal, construct, external | Yes No Deduced No No No No No No No No Partially No Partially No Partially | size, MRR type of token? Dataset Same as training test NVIDIA Tesla V100, 4 GPUs |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Yes No Deduced No No No No No No No Partially No Partially No Partially No Partially No | size, MRR type of token? Dataset Same as training test NVIDIA Tesla V100, 4 GPUs Overall/mean |

Paper Code prediction by feeding trees to tra

ID AP23 Experiments 4

Comments E4 is included in related work, at the ϵ

| Aspect | Element | E3 | Comments |
|---------------------|---|-------------------------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Madal parameters | Samo as provious | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | 224,60 | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| | | | |
| | | | |
| Operationalization | Factors and treatments | TravTrans, TravTrans+ (2 proposals, | _ |
| • | | the differences are inputs/outputs | |
| | | only) | |
| | | | |
| | Response variable, elaboration and metric | Yes | Training time (min/epoch), inference time, model |
| | | | size, MRR |
| Design | Design type | No | |
| | Blocking variables | Deduced | type of token? Dataset |
| | Held-constant variables | No | |
| | Measured variables (covariates) Randomization | No | |
| | Task duration | No No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training test |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | NVIDIA Tesla V100, 4 GPUs |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | Overall/mean |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Code prediction by feeding trees to tra

ID AP23 Experiments 4

Comments E4 is included in related work, at the ϵ

| Aspect | Element | E4 | Comments |
|---------------------|---|---------------------------|---------------------------|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| | | | |
| | | | |
| 0 " " " | | D | |
| Operationalization | Factors and treatments | PointerMixture, TravTrans | |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Partially | out-of-vocabulary rates |
| | | | |
| Design | Design type | No | |
| | Blocking variables | Deduced | type of token? Dataset |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure Number of experimental units | No | |
| Instrumentation | Test set | No Portially | Same as training test |
| msuumentauun | Measuring instruments | Partially No | Same as training test |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | NVIDIA Tesla V100, 4 GPUs |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | Overall/mean |
| , | Inferential statistics | No | • • |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

Paper A context-based automated approach for method name consistency checking and suggestion

ID AP24 Experiments 6

| Aspect | Element | E1 | Comments |
|-----------------------------|---|----------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | |
| | | #neurons/layer: No | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: 80-10-10 | Hyperparameters automatically fine-tuned |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Public dataset (used in a previous work) |
| Operationalization | Factors and treatments | Partially | Liu et al., Mnire |
| | Response variable, elaboration and metric | | Precision, recall, F-score (for both IC and C) |
| Design | Design type | No | |
| · · | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | Yes | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Malialia da da la castia da | Conclusion, internal, construct, external | Partially | very poor |
| validity evaluation | | | |
| Artifact | Availability | Yes | |

Paper A context-based automated approach

ID AP24 Experiments 6

| LAPELIIII | |
|-----------|----|
| Commen | ts |

| Aspect | Element | E2 | Comments |
|---------------------|------------------------|------------------|----------|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | · | |
| | | | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Yes | Public dataset (used in a previous work) |
|---------------------|---|-----------|--|
| Operationalization | Factors and treatments | Partially | Mnire, code2vec, code2seq, path-based |
| | | | representation |
| | Response variable, elaboration and metric | Yes | Precision, recall, F-score |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | Yes | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | very poor |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper A context-based automated approach

ID AP24 Experiments 6

| Com | me | nts |
|-----|----|-----|

| Aspect | Element | E3, E4 | Comments | |
|---------------------|--------------------------|------------------|-----------|--|
| Experiment type | | Optimization | Two tasks | |
| Hypotheses | Research hypotheses | Yes | RQ3 | |
| | Statistical hypotheses | No | | |
| Variables selection | on Model hyperparameters | Same as previous | | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Yes | Public dataset (used in a previous work) |
|---------------------|---|-----------|--|
| Operationalization | Factors and treatments | Partially | Model versions: Internal, enclosing, siblings, |
| | | | interaction |
| | Response variable, elaboration and metric | Yes | Exmatch, precision, recall, F-score |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | Yes | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | very poor |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper A context-based automated approach

ID AP24 Experiments 6

Comments

| Aspect | Element | E5, E6 | Comments |
|---------------------|------------------------|------------------|-----------|
| Experiment type | | Optimization | Two tasks |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Yes | Public dataset (used in a previous work) |
|---------------------|---|-----------|--|
| Operationalization | Factors and treatments | Partially | Model versions: Seqseq, seq2seq+copy, |
| | | | seq2sec+copy+non-copy |
| | Response variable, elaboration and metric | Yes | Exmatch, precision, recall, F-score |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | Yes | |
| | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | Yes | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | very poor |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper DeepSim: Deep learning code functional similarity

ID AP25 Experiments 3

Comments E1 is hidden

| Aspect | Element | E1 | Comments |
|---------------------|---|-----------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as next | |
| | Model parameters | Same as nest | |
| | DL algorithm | Same as next | |
| | Training hyperparameters | Same as next | |
| | Training data | Same as next | |
| Operationalization | Factors and treatments Response variable, elaboration and metric | No Partially | Mentions hyperparameters (e.g. learning rate, layer sizes, regularization rate, dropout rate, various acrivation functions and weights initializers, etc.) But does not say which ones. For all models compared Testing errors and F-score |
| Design | Design type | No V | Detect 10 feld over cell-felice |
| | Blocking variables | Yes | Dataset, 10-fold cross-validation For RtVNN same full dataset |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure Number of experimental units | No Yes | 10 (10-fold cross-validation) |
| Instrumentation | Test set Measuring instruments Measurement procedure Technological infrastructure | No No No Yes | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics Inferential statistics | No No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| | | | |
| Artifact | Availability | Yes | |

Paper DeepSim: Deep learning code function

ID AP25 Experiments 3

Comments E1 is hidden

| Aspect | Element | E2 | Comments |
|---------------------|---|-------------------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| (1 | Statistical hypotheses | No | For all we delected as a size |
| variables selection | Model hyperparameters | #layers: Not clear | For all models: layers size |
| | | #neurons/layer: 88-6, (128x6-256- | |
| | | 64)-128-32 | |
| | | activation functions: ELU | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | For all models, L2 value & dropout |
| | | model type: Yes | |
| | | loss function: cross-entropy | |
| | | regularization: L2(0.00003) | |
| | | optimization: dropout (0.75) | |
| | Training hyperparameters | train-test split: 10-fold-cross.val | For all models, learning rate and epochs |
| | | learning rate: 0.001 (initial) | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 4 | |
| | Training data | Yes | Perfect!!! Google Code Jam |
| Operationalization | Factors and treatments | Partially | Model: DECKARD (no-DNN, stable Github version), |
| | | | SdA-base, SdA-unsup |
| | | | RtvNN (NN, not available, re-implementation |
| | | | according to paper), |
| | | | DeepSim |
| | Response variable, elaboration and metric | Partially | Recall, precision, F1, time. Only mentions name. |
| | | | Training time includes generation of semantic feature |
| | | | matrices from bytecode files for DeepSim and two |
| | | | SdA baseline models. |
| Design | Design type | No | |
| | Blocking variables | Yes | Dataset, 10-fold cross-validation. Reported result is averaged |
| | Held-constant variables | No | · · |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | Yes | 10-fold cross-validation for DeepSim and SdA. For |
| | | | RtVNN same full dataset |
| | | | For DECKARD, 3 runs and report average |
| Instrumentation | Test set | Yes | Perfect!!! |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | WALA geneate semantic feature matrix. Tensorflow. |
| | | | Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU |
| Population | Objects (chars. of the experimental datasets) | Yes | Perfect!!! |
| Analysis | Descriptive statistics | Partially | Mean only (10-fold cross-validation and DECKARD) |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper DeepSim: Deep learning code function

ID AP25 Experiments 3

Comments E1 is hidden

| Aspect | Element | E3 | Comments |
|-------------------------------------|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | - " |
| Variables selection | Model hyperparameters | #layers: Not clear | For all models: layers size |
| | | #neurons/layer: 88-6, (128x6-256- | |
| | | 64)-128-32 | |
| | | activation functions: ELU | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | For all models, L2 value & dropout |
| | | model type: Yes | |
| | | loss function: cross-entropy | |
| | | regularization: L2(0.00003) | |
| | | optimization: dropout (0.75) | |
| | Training hyperparameters | train-test split: 10-fold-cross.val | For all models, learning rate and epochs |
| | | learning rate: 0.001 (initial) | , |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 4 | |
| | Training data | Yes | Perfect!!! BigCloneBench. Two trainings, full set and |
| | railing data | ies | set with functionality id=4 only |
| Operationalization | Factors and treatments | Partially | Model: DECKARD, RtvNN, CDLH (for the 3, values |
| Operationalization | ractors and treatments | raitally | reported in paper), DeepSim |
| | | | |
| | | | |
| | Response variable, elaboration and metric | Partially | Recall, precision, F1 and per clone type |
| Design | | | Recall, precision, F1 and per clone type |
| Design | Design type | No | |
| Design | | | For DeepSim. For the paper models, no n-fold cross |
| Design | Design type Blocking variables | No Partially | |
| Design | Design type Blocking variables Held-constant variables | No Partially | For DeepSim. For the paper models, no n-fold cross |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No Partially No No | For DeepSim. For the paper models, no n-fold cross |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | No Partially | For DeepSim. For the paper models, no n-fold cross |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No Partially No No No | For DeepSim. For the paper models, no n-fold cross |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No Partially No No No | For DeepSim. For the paper models, no n-fold cross validation |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No Partially No No No No No | For DeepSim. For the paper models, no n-fold cross validation |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No Partially No No No No No No Yes | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset |
| Design Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No Partially No No No No No Yes | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No Partially No No No No No Yes Yes Deduced | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No Partially No No No No No Yes Yes Deduced Deduced | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No Partially No No No No No Yes Yes Deduced | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No Partially No No No No No Yes Yes Deduced Deduced Partially | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No Partially No No No No No Yes Yes Deduced Deduced Partially Yes | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU Perfect!!! |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No Partially No No No No No Ves Yes Deduced Deduced Partially Yes Partially | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Partially No No No No No Yes Yes Deduced Deduced Partially Yes | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU Perfect!!! |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No Partially No No No No No No Yes Yes Deduced Deduced Partially Yes Partially No | For DeepSim. For the paper models, no n-fold cross validation 10-fold cross-validation for DeepSim. Reported result is averaged. For RtVNN same full dataset Perfect!!! WALA geneate semantic feature matrix. Tensorflow. Desktop PC with intel i7 4.0GHz 4 cores CPU, GTX 1080 GPU Perfect!!! |

ID AP26 Experiments 4

Comments E1 was hidden. Embedded in description of training

| Aspect | Element | E1 | Comments |
|---------------------|---|--|---|
| Experiment type | | Optimization | Totally hidden |
| | | | |
| Hypotheses | Research hypotheses Statistical hypotheses | No No | |
| Variables selection | Model hyperparameters | #layers: 7 (?) | Not clear n. Layers (missing from Figure 2 |
| | | #neurons/layer:n 300 | input and softmax) |
| | | (embedding), 650 (both bi-GRU), ? | Not clear if 2 bi-GRU layers, or they are the |
| | | (rest) | same |
| | | connections: ? (Input), embedding, | |
| | | bi-directional GRU, concatenation, | |
| | | bi-directional GRU, ? (projection), | |
| | | softmax, (?) output | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Model is selected when error is stabilized |
| | | model type: Yes | Momentum =1/e after the first 1,000 |
| | | loss function: cross-entropy | minibatches |
| | | regularization: dropout prob. of | |
| | | 50% (second hidden layer), layer normalization | |
| | | optimization: Adam, momentum | |
| | Training hypernarameters | • | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | | learning rate: 10 ⁻³ and reduce it | |
| | | every epoch until 10 ⁻⁴ | |
| | | #iterations: No | |
| | | batch size: 5,000 | |
| | Torining date | #epochs: 10 | 1 000 to a standard and a second a second and |
| | Training data | Partially??? | 1,000 top starred open-source projects on |
| | | | Github Feb 28,2018. Available. Not sure if |
| | | | reproducible. "Predominantly consisted of |
| Operationalization | Factors and treatments | No | TypeScript code" What is predominantly? Unknown which hyper-parameters were |
| Operationalization | ractors and treatments | NO | fine-tuned and how |
| | Response variable, elaboration and metric | Partially | Validation error |
| | • | · | |
| Design | Design type | No No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | Caaraa 1 mm |
| Instrumentation | Number of experimental units Test set | No Partially??? | Seems 1 run Same as training |
| mstrumentation | restact | Tortunity::: | Same as daming |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | Yes | Coded in CNTK. NVIDIA Geforce GTX 1080 Ti |
| | | | GPU with 11GB, 6-core Intel 17-8700 with |
| | | | 32GB RAM. Model needs 500MB RAM to be |
| | | | loaded into memory, and can be run on both |
| | | | GUP and CPU. Answer in well uncer 2 |
| B Lut. | Object the south | D. W. II | seconds |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics Inferential statistics | No No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

ID AP26 Experiments 4

Comments E1 was hidden. Embedded in descripti

| Aspect | Element | E2 | Comments |
|---------------------|------------------------|------------------|---|
| Experiment type | | Evaluation | Compares against naïve and plain RNN. It is not |
| | | | evaluation against SOTA. But the goal is not to |
| | | | "improve" but to "compare" |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | Seems the better option is chosen. |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

Training data Same as previous

| Operationalization | Factors and treatments | Partially | Naïve, plain rNN, DeepTyper |
|---------------------|---|--------------|---|
| | Response variable, elaboration and metric | Yes | Top-1, top-5 prediction accuracy at different types |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially??? | Same as training |
| | | | |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | Yes | Coded in CNTK. NVIDIA Geforce GTX 1080 Ti GPU |
| | | | with 11GB, 6-core Intel 17-8700 with 32GB RAM. |
| | | | Model needs 500MB RAM to be loaded into |
| | | | memory, and can be run on both GUP and CPU. |
| | | | Answer in well uncer 2 seconds |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP26 Experiments 4

Instrumentation Test set

Comments E1 was hidden. Embedded in descripti

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Evaluation | Compares DT, CJ and hybrid |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | Seems the better option is chosen. |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | DeepTyper, TypeScript compiler (non- |
| | Response variable, elaboration and metric | Yes | DNN), Hybrid accuracy (top-1), hitsm misses (for Hybrid |
| Design | Design type | No | |
| - | Blocking variables | No | Different "settings" are used (allowed-to- vary?) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| netrumentation | Tost set | No | Cama as provious? |

| | Measuring instruments | No | Can be deduced |
|---------------------|---|-----------|---|
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | Yes | Coded in CNTK. NVIDIA Geforce GTX |
| | | | 1080 Ti GPU with 11GB, 6-core Intel 17- |
| | | | 8700 with 32GB RAM. Model needs |
| | | | 500MB RAM to be loaded into memory |
| | | | and can be run on both GUP and CPU. |
| | | | Answer in well uncer 2 seconds |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

No

Same as previous?

ID AP26 Experiments 4

Comments E1 was hidden. Embedded in descripti

| Aspect | Element | E4 | Comments |
|---------------------|------------------------|------------------|------------------------------------|
| Experiment type | | Evaluation | Compares agains JSNice |
| | | | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | Seems the better option is chosen. |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

Training data Same as previous

| Operationalization | Factors and treatments | Partially | DeepTyper, JSNice (not DNN), Hybrid |
|---------------------|---|-----------|--|
| | Response variable, elaboration and metric | Partially | correct, partial, incorrect, unsure |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Is a different one. 30 randomly selected |
| | | | JavaSCript functions in GitHub projects |
| | | | that were in the top 100 projects ranked |
| | | | by numnber of stars |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | Yes | Coded in CNTK. NVIDIA Geforce GTX |
| | | | 1080 Ti GPU with 11GB, 6-core Intel 17- |
| | | | 8700 with 32GB RAM. Model needs |
| | | | 500MB RAM to be loaded into memory, |
| | | | and can be run on both GUP and CPU. |
| | | | Answer in well uncer 2 seconds |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics | No | |
| • | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Neural-augmented static analysis of android communication

ID AP27 Experiments 1

Comments

| Aspect | Element | E1 | Comments |
|---------------------|---|---|--|
| Experiment type | | Evaluation | Among alternatives designed by authors |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers:input+hidden (varies according to | - CNN: kernel sizes (1,3,5,7), kernel counts |
| | | instantiation)+output | (8,16,32,64), max pooling |
| | | #neurons/layer: Varies acording to | - RNN(LSTM): hidden size(128) |
| | | instantiation | - 1-layer perceptron (64) |
| | | connections: varies according to | - Multilayer perceptron (16,1) |
| | | instantiation | |
| | | activation functions: relu | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: cross-entropy | |
| | | regularization: No | |
| | | optimization: RMSProp | |
| | Training hyperparameters | train-test split: Yes | Training set: 105,108 links, testing set: |
| | | learning rate: No | 43,680 (may) links |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Partially | Only reference.PRIMO corpus |
| Operationalization | Factors and treatments | Partially | Instantiations (DNN model): str-RNN, str- |
| | | | CNN, typed-simple, syped-tree |
| | Response variable, elaboration and metric | Partially | No formulas. F1, AUC, Kurskal's gamma, |
| | | | #trainable parameters, inference time, |
| | | | entropy probability of true positives, portion |
| | | | of link with such high predicted values |
| Design | Design type | No | |
| _ | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | Partially | PRIMO |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Keras with Tensorflow and IC3 |
| | | | Intel Core i7-6700 (3.4GHz) CPU, 32GB |
| | | | memory, 1TB SSD, NVIDIA GeForce GTX 970 |
| - I .: | | | GPU (trained & tested on GPU) |
| Population | Objects (chars. of the experimental datasets) | No | PRIMO |
| Analysis | Descriptive statistics | No | |
| Validity ovalvation | Inferential statistics | No Partially | Elat list of 2 throats. No link to times |
| • | | Partially | Flat list of 2 threats. No link to types |
| Artifact | Availability Badge | No No | |
| | DAURE | INU | |

Paper Complementing global and local contexts in representing API descriptions to improve API retrieval tasks

ID AP28 Experiments 3

Comments Each experiment evaluates approach for a different task. Changes "on the fly" for datasets

| Aspect | Element | E1 | Comments |
|---------------------|---|---------------------------------|--|
| Experiment type | | Evaluation | text-to-code retrieval |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 3? | |
| | | #neurons/layer No | |
| | | connections: input, | |
| | | concatenate, classifier | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Partial | |
| | 224,6014 | model type: No | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Tarining by several several | • | |
| | Training hyperparameters | train-test split: are different | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Partial | API documentation of Java JDK core library |
| | | | [6,21] |
| Operationalization | Factors and treatments | Partial | D2Vec, Word2Vec (both with C= 5, N=200), |
| | | | rVSM (no DNN), rVSM+Word2Vec, |
| | | | rVSM+D2Vec |
| | Response variable, elaboration and metric | Yes | top-k accuracy [15] |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partial | Java tutorial website (KodeJava) |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Partial | Some statistics are given (table 1) |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partial | Plain list and very weak |
| Artifact | Availability | No | |
| | Badge | No | |

Paper Complementing global and local conte

ID AP28 Experiments 3

Comments Each experiment evaluates approach 1

| Aspect | Element | E2 | Comments |
|---------------------|---|---------------------------------------|---|
| Experiment type | | Evaluation | code-tot-text retrieval |
| Hypotheses | Research hypotheses | yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | · · | • | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | Except train-test split |
| | | , , , , , , , , , , , , , , , , , , , | |
| | | | |
| | | | |
| | | | |
| | Training data | Partial | Same as test. Same used for FRAFT |
| | Training data | ratuat | Same as test. Same asea for Fixer 1 |
| Operationalization | Factors and treatments | Partial | FRAFT, Word2Vec, D2Vec |
| o perationalization | ractors and acadiments | rardar | 11011 1, 000102000, 020000 |
| | | | |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | No | 5 libraries. Not sure what they are |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partial | The same used for FRAFt. Merely references it |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| Donulation | Technological infrastructure | No Partial | Comp statistics are siven (table 6) |
| Population Applysis | Objects (chars. of the experimental datasets) Descriptive statistics | Partial Partial | Some statistics are given (table 6) |
| Analysis | Inferential statistics | No | Only mean |
| Validity evaluation | | Partial | Plain list and very weak |
| Artifact | Availability | No | riaiii iist aliu vely weak |
| Aithact | Badge | No | |
| | buuge | 110 | |

Paper Complementing global and local conte

ID AP28 Experiments 3

Comments Each experiment evaluates approach 1

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Evaluation | code-to-code retrieval. Assessment |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | Not sure if there are 2 trainings or 1 |
| | Training hyperparameters | Same as previous | Except train-test split |
| | Training data | Partial | API documentation of Java JDK core and |
| | | | Apache Commons libraries [6,21] |
| Operationalization | Factors and treatments | Partial | D2Vec, Word2Vec |
| | Response variable, elaboration and metric | Yes | Top-k accuracy (1,2,3,5,10) |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Oracle API mappings provided in [38] |
| | Measuring instruments | No | Can be deduced |
| | Measurement procedure | No | Can be deduced |
| | Technological infrastructure | No | - |
| Population | Objects (chars. of the experimental datasets) | Partial | Same as E1 |
| Analysis | Descriptive statistics | No | Same as E1 |
| , 0.0 | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Plain list and very weak |
| Artifact | Availability | No | . Jan. Hot arta tory troat |
| Artifact | | | |

Paper On using machine learning to identify knowldge in API reference documentation

ID AP29 Experiments 1

Comments

| Aspect | Element | E1 | Comments |
|---------------------|---|--|---|
| Experiment type | | Evaluation | SOTA |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 5 | |
| | | #neurons/layer: 300, ?, 128, 64, 12 | |
| | | connections: input, LSTM, Dense, | |
| | | Dense, output | |
| | | activation functions: tahn,, ReLU, | |
| | | ReLU, sigmoid | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Input word embedding vectors trained using |
| | | model type: Yes | GloVe |
| | | loss function: sigmoidal cross- | Siove |
| | | _ | |
| | | entropy | |
| | | regularization: No | |
| | | optimization: dropout (en LSTM), | |
| | | Adam | |
| | Training hyperparameters | train-test split: Yes | 10-fold cross-validation using 10% of |
| | | learning rate: 0.001 | dataset as test set |
| | | #iterations: No | |
| | | batch size: 32 | |
| | | #epochs: 100 | |
| | Training data | Yes (paper mentions code and | CADO. Resampling is made to improve it |
| | | data of the study are shared | |
| Operationalization | Factors and treatments | Partially (at different levels) | Two algorithms (k-NN and SV), and RNN with LSTM layer architecture, naïve (MF1, MF2, RAND)) |
| | Response variable, elaboration and metric | Partially | Not all are described at the same level of detail |
| | | | AUPRC (per knowledge type), hamming loss, subset accuracy, macroprecision, macrorecall, macroF1, macroAUC |
| Design | Design type | No | |
| | Blocking variables | Corpora used to train embeddings Knowledge type | Glove is trained on 4 corpora for RNN |
| | | Test set | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | 10 | 10-fold cross-validation using 10% of dataset as test set |
| Instrumentation | Test set | Partially | New Python dataset (not very well explained) |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | GloVe for embeddings, trained on 4 corpora |
| Population | Objects (chars. of the experimental datasets) | Partially | New Python dataset (not very well explained) |
| Analysis | Descriptive statistics | Partially | 10-fold cross-validation, assume using means |
| • | Inferential statistics | No | , 3 |
| Validity evaluation | | Partially | Not linked in al cases to classification |
| Artifact | Availability | Yes | |
| | Badge | Yes | |
| | <u> </u> | | |

Paper Maximal multi-layer specification synthesis

ID AP30 Experiments 3

Comments E1 is hidden. Grid search for hyperparameters

| Aspect | Element | E1 | Comments |
|------------------------------|---|--|--|
| Experiment type | | Optimization | Grid Search used |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 4? | |
| | | #neurons/layer: No | |
| | | connections: input, encoder (LSTM), | |
| | | decoder (LSTM), output | |
| | | activation functions: encoder=sigmoid, | |
| | | tanh | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: negative log likelihood | |
| | | regularization:No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: No | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Partially | 20,640 pages from Stackoverflow using the |
| | | | search keywords "tidyr" and "dplyr",, |
| | | | removing duplicates and questions with no |
| | | | solutions |
| Operationalization | Factors and treatments | Number of neurons of the word/function | Grid search is used |
| | | embedding layer and LSTM hidden layer | |
| | Response variable, elaboration and metric | No | |
| | nesponse variable, claboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No Double It | Determine from a consideration of the constant |
| | Technological infrastructure | Partially | Pytorch framework. Google cloud platform |
| | | | with 2.20GHz intel xeon and NVIDIA Tesla |
| | | | K80 GPU |
| | | | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| Validity over treet | Inferential statistics | No Double II. | 2 threats not estador |
| Validity evaluation Artifact | | Partially No | 2 threats, not categorized |
| Aidldl | Availability Badge | No No | |
| | Dauge | INO | |

Paper Maximal multi-layer specification synthesis

ID AP30 Experiments

3 E1 is hidden. Grid search for hyperparameters Comments

| Aspect | Element | E2 | Comments |
|------------------------|--|-------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | The second secon | | |
| | DL algorithm | Same as previous | |
| | - | • | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| | | | |
| Operationalization | Factors and treatments | Refers to architecture | multi-layer spec and neural architecture (n-gram, |
| Operationalization | ractors and deatherns | nerers to are intecture | seq2seq, hybrid) |
| | | | 304 <u>2304</u>), 2 |
| | Response variable, elaboration and metric | Yes | Ranking of the correct candidate that matches |
| | | | the user intent. Counts of top-1s and top-3s |
| Design | Design type | No | |
| | Blocking variables | 2 libraries | But no separate results are given |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration Procedure | No No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Original benchmarks from Morpheus |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Pytorch framework. Google cloud platform with |
| | | | 2.20GHz Intel Xeon and NVIDIA Tesla K80 GPU |
| | | | |
| | | | |
| | | | |
| Danielation | Objects (share of the considerated data) | NI- | Defenses to seem |
| Population Analysis | Objects (chars. of the experimental datasets) | No Yes | Reference to paper |
| MIIdiySiS | Descriptive statistics Inferential statistics | Yes No | Mean and std. Dev for ranking |
| Validity evaluation | | Partially | 2 threats, not categorized |
| Artifact | Availability | No | |
| | Badge | No | |
| | | | |

Paper Maximal multi-layer specification synthesis

ID AP30 Experiments 3

Comments E1 is hidden. Grid search for hyperparameters

| Aspect | Element | E3 | Comments |
|---------------------|---|-------------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| Variables selection | Statistical hypotheses Model hypotheses | No Different from provious | word/function embedding layer and LCTM |
| variables selection | Model hyperparameters | Different from previous | word/function embedding layer and LSTM hidden layer to be 256, where the embedding layer maps 25,004 words and 14 functions to vectors of the dimension 256. A single layer perceptron is connected to the hidden layer of each output time step in the decoder, mapping from a dimension of 512 (256x2) to 14 |
| | Model parameters | biases: No | 110111 d diffiction of 512 (250x2) to 14 |
| | Woder parameters | weights:No | |
| | DI algorithm | | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Missing | Same as previous??? |
| Operationalization | Factors and treatments Response variable, elaboration and metric | Refers to architecture Yes | multi-layer spec and neural architecture (n-gram, seq2seq, hybrid) Time |
| | | | |
| Design | Design type | No | |
| | Blocking variables | 2 libraries | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | Yes | Limit 300 secs |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Original benchmarks from Morpheus |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | All synthesis tasks were run on a laptop |
| | | | equipped with Intel Core i5 CPU and 16GB |
| | | | memory. Since the Morpheus tool is only |
| | | | available on a virtual machine, we used this |
| | | | virutal machine to run all program synthesis |
| | | | experiments. |
| Population | Objects (chars. of the experimental datasets) | No | Reference to paper |
| Analysis | Descriptive statistics | Partially | mean |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | 2 threats, not categorized |
| Artifact | Availability | No | |
| | Badge | No | |

Paper

Robust log-based anomaly detection on unstable log data

ID

AP31 2

Experiments Comments

| Aspect | Element | E1 | Comments |
|---------------------|---|---|--|
| Experiment type | | Evaluation | SOTA |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 4 | |
| | | #neurons/layer: | |
| | | connections: input, bi-LSTM, fully, output | |
| | | activation functions: fully=tanh, | |
| | | ouput=softmax | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | Dealgonami | model type: Yes | |
| | | * * | |
| | | loss function: cross-entropy | |
| | | regularization: weight decay (L2)=0.0001 | |
| | | optimization: momentum=0.9 | |
| | Training hyperparameters | train-test split: No | |
| | | learning rate: No | |
| | | #iterations: No | |
| | | batch size: 128 | |
| | | #epochs: 10 | |
| | Training data | Partially (see test set) | Training Synthetic HDFS datset |
| Operationalization | Factors and treatments | DNN (LogRobust, SVM, LR, IM, PCA) | The rest are non-DNNs |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | Injection ratio | Not applicable for stable HDFS dataset |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially. Collect randomly. Refer to another | Test sets synthetic HDFS dataset |
| | | paper for HDFS. No info about MS due to | (NewTesting1, NewTesting2), Microsoft |
| | | confidentiality | industrial dataset, stable HDFS dataset |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Keras, NVIDIA Tesla M40 GPU |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Separate "analyses" for different datset |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | No | |
| | Badge | No | |

Paper Robust log-based anomaly detection c

ID AP31 Experiments 2

Comments

| Aspect | Element | E2 | Comments | |
|-----------------|------------------------|--------------|----------|--|
| Experiment type | 1 | Optimization | | |
| Hypotheses | Research hypotheses | Yes | RQ2 | |
| | Statistical hypotheses | No | | |
| | | | | |

Variables selection Model hyperparameters Same as before

Model parameters Same as before

DL algorithm Same as before

Training hyperparameters Same as before

| | Training data | Same as before | |
|---------------------|---|---------------------------------------|-----------------------------------|
| Operationalization | Factors and treatments | Architecture (with/without attention) | |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | Injection ratio | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Same as before | NewTesting2 |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Keras, NVIDIA Tesla M40 GPU |
| Population | Objects (chars. of the experimental datasets) | Yes | nerasji w i zavi i esia w i e e e |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | No | |
| | Badge | No | |

Paper SEntiMoji: An emoji-powered learning approach for sentiment analysis in SE

ID AP32 Experiments 2

Comments The proposal (SEntiMoji) is built upon DeepMoji, and then fine-tuned on a different task

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | skip-gram for word embeddings |
| | | #neurons/layer: No | Replace 64-dimension softmax layer of |
| | | connections: Yes | DeepMoji with an n-dimension softmax |
| | | activation functions: No | layer |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: No | DNN is pre-trained in a different task, later |
| | 9 7 | learning rate: No | fine-tuned. They build SEntiMoji upon |
| | | #iterations: No | DeepMoji |
| | | batch size: No | . , |
| | | #epochs: No | |
| | Training data | Yes | Linked to repo |
| Operationalization | Factors and treatments | Partially | Senti-Strength, SentiStrength-SE, SentiCR, Senti4SD, SentiMoji (the others are not DNNs) |
| | Response variable, elaboration and metric | Partially | Precision, recall, F-score, accuracy |
| Design | Design type | No | · · · · · · · |
| 0 | Blocking variables | Deduced | 5-fold-cross-validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | Partially | Averaged values |
| | Inferential statistics | Yes | McNemar test |
| Validity evaluation | | Partially | Construct, internal, external |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper SEntiMoji: An emoji-powered learning

ID AP32 Experiments 2

Comments The proposal (SEntiMoji) is built upon

| Aspect | Element | E2 | Comments |
|---------------------|---|------------------|--------------------------------------|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| O | Training data | Yes | Linked to repo |
| Operationalization | Factors and treatments | Partially | SentiMoji, SentiMoji-G, SentiMoji-T |
| | Response variable, elaboration and metric | Partially | Precision, recall, F-score, accuracy |
| Design | Design type | No | |
| | Blocking variables | Deduced | 5-fold-cross-validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | Partially | Averaged values |
| | Inferential statistics | Yes | McNemar test |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Construct, internal, external |
| Artifact | Availability | Yes | • • • |
| | Badge | No | |
| | <u> </u> | | |

Paper SEntiMoji: An emoji-powered learning

ID AP32 Experiments 2

Comments The proposal (SEntiMoji) is built upon

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--------------------------------------|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| /ariables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Linked to repo |
| nerationalization | Factors and treatments | Partially | Train test size |
| perationalization | ractors and deatherts | rardany | 110111 test size |
| | | | |
| | Response variable, elaboration and metric | Partially | Precision, recall, F-score, accuracy |
| esign | Design type | No | |
| | Blocking variables | Deduced | 5-fold-cross-validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| nstrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| opulation | Objects (chars. of the experimental datasets) | Yes | Described |
| nalysis | Descriptive statistics | Partially | Averaged values |
| | Inferential statistics | Yes | McNemartest |
| alidity evaluation | Conclusion, internal, construct, external | Partially | Construct, internal, external |
| rtifact | Availability | Yes | |
| | Badge | No | |

Paper TypeWriter: natural type prediction with search -based validation

ID AP33 Experiments 3

Comments One more experiment about issues outside DNN

| Variables selection Model Model DL Tra Operationalization Face | esearch hypotheses tatistical hypotheses Todel hyperparameters Todel parameters L algorithm | Evaluation Yes No #layers: #neurons/layer: 200LSTM connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy regularization: No | SOTA RQ1 Word2vec for embeddings. Trained 2 times: code+identifiers and comments |
|--|---|--|--|
| Variables selection Model Model DL Tra Operationalization Face | tatistical hypotheses Todel hyperparameters Todel parameters L algorithm | No #layers: #neurons/layer: 200LSTM connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | Word2vec for embeddings. Trained 2 times: |
| Variables selection Model Model DL Tra Operationalization Face | Todel hyperparameters Todel parameters Lalgorithm | #layers: #neurons/layer: 200LSTM connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | _ |
| DL Tra Tra Operationalization Fac | Iodel parameters Lalgorithm | #neurons/layer: 200LSTM connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | - |
| Tra Operationalization Fac | L algorithm | connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | code+identifiers and comments |
| Tra Operationalization Fac | L algorithm | RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | layer activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | activation functions: ouput layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | layer (softmax) params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | params. Initialization: No biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | biases: No weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | L algorithm | weights: No representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | | representation: Yes model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | | model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | | model type: Yes loss function: cross-entropy | |
| Tra Operationalization Fac | raining hyperparameters | loss function: cross-entropy | |
| Tra Operationalization Fac | raining hyperparameters | • • | |
| Tra Operationalization Fac | raining hyperparameters | regularization. No | |
| Tra Operationalization Fac | raining hyperparameters | = | |
| Tra Operationalization Fac | raining hyperparameters | optimization: Adam | The entire neural model is learned is inthe |
| Operationalization Fac | . O .) harkaramana | train-test split: 80-20 | The entire neural model is learned jointly |
| Operationalization Fac | | learning rate: No | (word2vec?) |
| Operationalization Fac | | #iterations: No | 2 separate trainings fr argument types and |
| Operationalization Fac | | batch size 0.005 | function types |
| Operationalization Fac | | #epochs: 10 | Split is by file |
| · | raining data | Partially | One of the 2 datasets is private (Facebook). |
| · | | | Other available for download |
| Res | actors and treatments | Partially | Model: naïve, DeepTyper (re- implementation), NL2Type (re- |
| Res | | | implementation) |
| | esponse variable, elaboration and metric | Yes | Precision, recall, F1 (weighted), top-k scores (1,3,5) |
| Design Des | esign type | No | (1,3,3) |
| - | locking variables | Datasets | 2: public and private |
| | eld-constant variables | No | 2. public and private |
| | leasured variables (covariates) | No | |
| | andomization | No | |
| | ask duration | No | |
| | rocedure | No | |
| | umber of experimental units | No | Seems 1 run |
| | est set | Partially | One of the 2 datasets is private (Facebook). |
| | | , | Other available for download |
| Dro | reprocessing | Yes | Word2vec using gensim, LibCST |
| | leasuring instruments | Deduced | TO SERVE GOING BETOILT, LIDEO I |
| | leasurement procedure | Deduced | |
| | echnological infrastructure | Partially | Python |
| | bjects (chars. of the experimental datasets) | Partially | For public dataset |
| | escriptive statistics | No | • |
| • | of a statistics | No | |
| | 11C1C11U013U03U03 | No | |
| • | | Partially | Independent re-implementation. La original es de Facebook |
| Вас | onclusion, internal, construct, external vailability | | |

Paper TypeWriter: natural type prediction w

ID AP33 Experiments 3

Comments One more experiment about issues ou

| Aspect | Element | E2 | Comments |
|--|--|--|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Same as previous Partially | Model: TypeWriter variants-removing parts (ful |
| ореганопанданоп | | | typemask, tokenseqs, names, documentation), NL2Type |
| | Response variable, elaboration and metric | Yes | Precision, Recall, 2 prediction levels |
| Design | Design type | No | |
| _ | Blocking variables | Datasets | 2: public and private |
| | Held-constant variables | No | · |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | | | |
| Instrumentation | Test set | Partially | One of the 2 datasets is private (Facebook). Other available for download |
| Instrumentation | Test set Preprocessing | Partially Yes | Other available for download |
| Instrumentation | | | |
| Instrumentation | Preprocessing | Yes | Other available for download |
| Instrumentation | Preprocessing Measuring instruments | Yes Deduced | Other available for download |
| | Preprocessing Measuring instruments Measurement procedure | Yes Deduced Deduced | Other available for download Word2vec using gensim, LibCST |
| Population | Preprocessing Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Yes Deduced Deduced Partially | Other available for download Word2vec using gensim, LibCST Python |
| Population | Preprocessing Measuring instruments Measurement procedure Technological infrastructure | Yes Deduced Deduced Partially Partially | Other available for download Word2vec using gensim, LibCST Python |
| Population Analysis | Preprocessing Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Yes Deduced Deduced Partially Partially No | Other available for download Word2vec using gensim, LibCST Python |
| Population Analysis Validity evaluation Artifact | Preprocessing Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Yes Deduced Deduced Partially Partially No | Other available for download Word2vec using gensim, LibCST Python |

Paper TypeWriter: natural type prediction wi

ID AP33 Experiments 3

Comments One more experiment about issues ou

| Aspect | Element | E3 | Comments |
|---------------------|--|------------------|-------------------------------------|
| Experiment type | | Evaluation | No-DNN (static type inference) |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | No | Internal dataset only |
| Operationalization | Factors and treatments | Partially | Model: TypeWriter, pyre |
| | Response variable, elaboration and metric | Yes | Added annotations vs. top-5 resutls |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | No | Private only |
| | Preprocessing | Yes | Word2vec using gensim, LibCST |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Python |
| Population | Objects (chars. of the experimental datasets) | No | Private dataset only |
| | Descriptive statistics | No | , |
| Analysis | Inferential statistics | No | |
| Allalysis | | | |
| | Conclusion, internal, construct, external | No | |
| | Conclusion, internal, construct, external Availability | No Partially | Independent re-implementation |

Paper Idnetifying linked incidents in large-scale online serice systems

ID AP34 Experiments 2

Comments other 2 are about variables outside DNN, one more is case study

| Aspect | Element | E1 | Comments |
|---------------------|---|--------------------------------------|---|
| Experiment type | | Evaluation/optimization | Compares against SOTA and variations of proposal |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | For embeddings, FastText and node2vec |
| | | #neurons/layer: No | |
| | | connections: convolutional+max-over- | Convolutional layer uses 3 sets of convolution |
| | | time pooling layer+fully connected | kernels with 3 different widhts (3,4,5), with 100 |
| | | activation functions: No | kernels each |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Training set: incidents 01/01/17 to 08/31/2018. |
| | 0 / - - - - - - - - - | learning rate: No | Testing set: incidents 09/01/2018 to 10/31/2018. |
| | | #iterations: No | Validation test is 5% training set |
| | | batch size: No | vandation test is 575 training set |
| | | #epochs: 30 | |
| | Training data | No | Privacy issues (Microsoft) |
| Operationalization | Factors and treatments | Partially | Methods: DWEN (DNN), DBTM, simple (noDNNS), |
| - | | | LIDAR-T, LiDAR-C, LIDAR (proposed) |
| | Response variable, elaboration and metric | Partially | Precision, recall, F1 (just names) |
| Design | Design type | No | recession, recan, r 1 quaeriames, |
| | Blocking variables | Yes | 10 applications |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | |
| | | | |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Ubuntu 16.04, 24-core dual-intel xeon E5-2690 V3 |
| | | | CPU (2.60GHz), 220 GB memory, and a single |
| | | | NVIDIA Testla K80 GPU |
| Population | Objects (chars. of the experimental datasets) | Partially | Cannot be disclosed |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | | No | |
| Artifact | Availability | No | |
| | Badge | No | |

Paper Idnetifying linked incidents in large-sc

ID AP34 Experiments 2

Comments other 2 are about variables outside DN

| Aspect | Element | E2 | Comments |
|---------------------|---|----------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | _ |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Testatos data | 6 | Drivers (address ft) |
| Operationalization | Training data Factors and treatments | Same as previous Partially | Privacy issues (Microsoft) Methods: DWEN (DNN), LIDAR (proposed) |
| Operationalization | ractors and deatments | raitany | Wethous. DWEN (DNN), LIDAN (proposed) |
| | Response variable, elaboration and metric | Partially | Precision |
| Design | Design type | No | |
| · · | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | 10,000 human-machine links and 10,000 |
| | | | machine-machine links randomly selected |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | Libraria 10 04 24 save dual intelluses EF 2000 |
| | Technological infrastructure | Partially | Ubuntu 16.04, 24-core dual-intel xeon E5-2690 |
| | | | V3 CPU (2.60GHz), 220 GB memory, and a single |
| Danulation | Objects (shows of the superior outs) | De stielle. | NVIDIA Testla K80 GPU |
| Population | Objects (chars. of the experimental datasets) | Partially | Cannot be disclosed |
| Analysis | Descriptive statistics Inferential statistics | No No | |
| Validity evaluation | | No | |
| Artifact | Availability | No | |
| | Badge | No | |
| | 0- | | |

Paper On the naturalness of hardware descriptions

ID AP35 Experiments 2

Comments El E1 está hidden. Habla de tuning pero no dice nada

| Aspect | Element | E1 | Comments |
|---------------------|--------------------------|-------------------------------|--|
| Experiment type | | Optimization | Hidden |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | |
| | | #neurons/layer: No | |
| | | connections: 2-6 encoders, | |
| | | fully connected, decoders | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Early stop |
| | | model type: bidirectional GRU | |
| | | loss function: No | |
| | | regularization: ReLU on fully | |
| | | connected layer | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | 5 / | learning rate: No | |
| | | #iterations: No | |
| | | batch size: 32 | |
| | | #epochs: No | |
| | Training data | Yes | Aunque no sé si repetible |
| Operationalization | Factors and treatments | No | Parece que hace el tuning para todas las DNNs que usa en el E2. |

| | Response variable, elaboration and metric | No | |
|---------------------|---|-----------|------------------|
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch, OpenNMT |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | PArtially | |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

Paper On the naturalness of hardware descr

ID AP35 Experiments 2

Comments El E1 está hidden. Habla de tuning per

| Aspect | Element | E2 | Comments |
|------------------------------|---|--------------------------------|--|
| Experiment type | | Optimization/evaluation | Parece combinación porque prueba distintas |
| | | | opciones |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 6(?) | |
| | | #neurons/layer: | |
| | | encoders/decoders 512 | |
| | | connections: 2 encoders, fully | |
| | | connected, 2 decoders | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Early stop |
| | | model type: bidirectional GRU | |
| | | loss function: No | |
| | | regularization: ReLU on fully | |
| | | connected layer | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | g, perparameters | learning rate: No | |
| | | #iterations: No | |
| | | batch size: 32 | |
| | | | |
| | Training data | #epochs: No | Aunque no sé si repetible |
| Operationalization | Training data Factors and treatments | Yes DNN model | Los treatments no parecen SOTA. Rule-based |
| Operationalization | ractors and deadners | DININTITIONET | baseline, language model baseline (RNNLM, |
| | | | |
| | | | RNNLM+PA(1),RNNLM+PA(1-5)), sequence-to- |
| | | | sequence (\$2\$, \$2\$+PA(1), \$2\$+PA(1)+Type, |
| | | | S2S+PA(1-2)+Type, S2S+PA(1-3)+Type, |
| | | | S2S+PA(1-4)+Type, S2S+PA(1-5)+Type, |
| | | | S2S+PA(Ensemb-1-5)+Type |
| | | | La diferencia con los RNN models es que single- |
| | | | directional GRU |
| | Response variable, elaboration and metric | Yes | BLEU, Accuracy, exact-match accuracy |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | 2 (random cood different training (validation (test cali |
| Instrumentation | Number of experimental units Test set | Yes Yes | 3 (random seed, different training/validation/test split |
| Instrumentation | Measuring instruments | Yes | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Pytorch, OpenNMT |
| Population | Objects (chars. of the experimental datasets) | Yes | . 1 o pointing |
| Analysis | Descriptive statistics | Partially | Solo mean |
| , 5.5 | Inferential statistics | Yes | |
| | | Partially | - |
| Validity evaluation | Conclusion, internal, construct, external | | |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Yes | _ |

Paper MTFuzz: Fuzzing with a multi-task neural network

ID Experiments Comments

AP36 4

| Element | E1 | Comments |
|--|---|---|
| | Evaluation | |
| Research hypotheses | Yes | RQ1 |
| Statistical hypotheses | No | |
| Model hyperparameters | #layers: 7 + 3(with 3 paralells) | Each task has the same weight |
| | #neurons/layer: L1(?), | |
| | L2(2048), L3(1024), L\$(512) | |
| | connections: 3 encoder, 3 (x3) | |
| | | |
| | | |
| | | |
| | = ' | |
| Model narameters | • | |
| Wodel parameters | | |
| DI alaasishaa | • | Land from the NACE for a deal and a second |
| DL algorithm | · | Loss function: MSE for edge coverage, |
| | * * | adaptive loss for edge and context-sensitive |
| | | edge |
| | regularization: No | |
| | optimization: Adam | |
| Training hyperparameters | train-test split: No | 750 input samples for re-training |
| | learning rate: 0.001 | |
| | #iterations: No | |
| | batch size: No | |
| | #epochs: 100 | |
| Training data | Yes | |
| Factors and treatments | Partially | Fuzzer: AFL, AFLFasst, FairFuzz, Angora (non- |
| | | DNNs), Neuzz(DNN), MTFuzz |
| | | |
| Response variable, elaboration and metric | Yes | Number of bugs detected, edge coverage |
| Design type | No | |
| Blocking variables | Program | 10 programs |
| Held-constant variables | No | |
| Measured variables (covariates) | No | |
| Randomization | No | |
| Task duration | Yes | 24 hours for real-world, 5 hours for synthetic |
| | | bugs |
| Procedure | No | |
| Number of experimental units | Yes | 5 repetitions to cover fuzzer variability |
| Test set | Yes | 2 datasets, one for real bugs, other for synthetic |
| Measuring instruments | Deduced | |
| Measurement procedure | Deduced | |
| Technological infrastructure | Yes | Keras 2.2.3 with Tensorflow-1.8.0. |
| | | Ubuntu18.04, Intel Xeon E5-2623, NVIDIA GTX |
| | | 1080Ti GPU. |
| | | For data collection, single core machine for an |
| | | |
| | | hour |
| Objects (chars, of the experimental datasets) | Partially | |
| Objects (chars. of the experimental datasets) Descriptive statistics | Partially Partially | Nothing for synthetic bugs |
| Descriptive statistics | Partially | |
| Descriptive statistics Inferential statistics | Partially No | Nothing for synthetic bugs For edge coverage mean and std. Dev. |
| Descriptive statistics | Partially | Nothing for synthetic bugs |
| | Research hypotheses Statistical hypotheses Model hyperparameters Model parameters DL algorithm Training hyperparameters Training data Factors and treatments Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | Research hypotheses Statistical hypotheses No Model hyperparameters #layers: 7 + 3(with 3 paralells) #neurons/layer: L1(?), L2(2048), L3(1024), L5(512) connections: 3 encoder, 3 (x3) decoder activation functions: ReLu for hidden, sigmoid output params. Initialization: biases: No weights: No DL algorithm representation: Yes model type: Yes loss function: multi-task regularization: No optimization: Adam train-test split: No learning rate: 0.001 #iterations: No batch size: No #epochs: 100 Training data Yes Factors and treatments Partially Response variable, elaboration and metric Design type Blocking variables Measured variables (covariates) No Randomization No Task duration Procedure No Number of experimental units Yes Measurement procedure Deduced Measurement procedure Deduced |

Paper MTFuzz: Fuzzing with a multi-task nei

ID AP36 Experiments 4

Comments

| Aspect | Element | E2 | Comments |
|---------------------|---|------------------|---|
| Experiment type | | Optimization | With some/without auxiliary tasks |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | Removing some of the decoders. All configs. Use |
| | | | same hyperparams, etc. |
| | | | |
| | Response variable, elaboration and metric | Yes | Edge coverage |
| Design | Design type | No | |
| | Blocking variables | Program | |
| | Held-constant variables Measured variables (covariates) | No No | |
| | Randomization | No | |
| | Task duration | Yes | 1 hour |
| | | | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Real bugs only |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Yes | Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu18.04, |
| | | | Intel Xeon E5-2623, NVIDIA GTX 1080Ti GPU. |
| | | | For data collection, single core machine for an |
| | | | hour |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics | No | |
| , | Inferential statistics | No | |
| Validity evaluation | | Partially | 3 threats not classified |
| Artifact | Availability | Yes | |
| | Badge | Yes | |
| | | | |

Paper MTFuzz: Fuzzing with a multi-task net

ID AP36 Experiments 4

Comments

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Optimization | Adaptive loss |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | Adaptive loss |
| | | | |
| | | | - " - |
| | Response variable, elaboration and metric | No | Recall, F1 |
| Design | Design type | No Dragger | |
| | Blocking variables | Program | |
| | Held-constant variables | No | |
| | Measured variables (covariates) Randomization | No No | |
| | Task duration | No | Not specified this time |
| | lask dulation | NO | Not specified this time |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Real bugs only |
| mstumentation | Measuring instruments | Deduced | near bugs offiy |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Yes | Keras 2.2.3 with Tensorflow-1.8.0. |
| | | | Ubuntu18.04, Intel Xeon E5-2623, |
| | | | NVIDIA GTX 1080Ti GPU. |
| | | | For data collection, single core machine |
| | | | for an hour |
| Population | Objects (chars. of the experimental datasets) | Partially | ioi airrioui |
| Analysis | Descriptive statistics | No | |
| Allulysis | Inferential statistics | No | |
| Validity evaluation | | Partially | 3 threats not classified |
| Artifact | Availability | Yes | o ancata not diagriffed |
| | Badge | Yes | |
| | • | | |

Paper MTFuzz: Fuzzing with a multi-task nei

ID AP36 Experiments 4

Comments

| Aspect | Element | E4 | Comments | |
|--------------------|-------------------------|------------------|----------|--|
| Experiment type | | Generalization | | |
| Hypotheses | Research hypotheses | Yes | RQ4 | |
| | Statistical hypotheses | No | | |
| Variables selectio | n Model hyperparameters | Same as previous | | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

| | Training data | Same as previous | |
|---------------------|---|------------------|--|
| Operationalization | Factors and treatments | Yes | Program type: ELF files, XML files, Fuzze (Neuzz, MTFuzz, AFL, MTFuzz |
| | | | inputs+embeddings) |
| | Response variable, elaboration and metric | Yes | Edge coverage |
| Design | Design type | No | 2050 0010.050 |
| 0 | Blocking variables | Program | 3 ELF, 2 XML |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | Yes | 1 hour |
| | Procedure | | |
| | Number of experimental units | | Seems 1 run |
| Instrumentation | Test set | Partially | Only reference |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Yes | Keras 2.2.3 with Tensorflow-1.8.0. |
| | | | Ubuntu18.04, Intel Xeon E5-2623, |
| | | | NVIDIA GTX 1080Ti GPU. |
| | | | For data collection, single core machine |
| | | | for an hour |
| Population | Objects (chars. of the experimental datasets) | Partially | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | 3 threats not classified |
| Artifact | Availability | Yes | |
| | Badge | Yes | |

Paper Automated constrution of energy test oracles for android

ID AP37 Experiments 4

Comments E1 hidden (grid search)

| Aspect | Element | E1 | Comments |
|---------------------|---|--------------------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: No | |
| | | #neurons/layer: No | |
| | | connections: No | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: weighted-cross | |
| | | entropy | |
| | | regularization: early stopping | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: No | |
| | 5 // 1 | learning rate: No | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Labeled dataset |
| Operationalization | Factors and treatments | Hyperparameters | Does not mention which ones |
| | Response variable, elaboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch. Laptop 2.2.GHz intel core i7 and 16GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Automated constrution of energy test

ID AP37 Experiments 4

Comments E1 hidden (grid search)

| Aspect | Element | E2 | Comments |
|---------------------|---|-------------------------|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ2, RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Partially | train-test split: 10-fold cross-val. |
| | | | |
| | | | |
| | | | |
| | Tueining data | Cama as annuitana | |
| Operationalization | Training data Factors and treatments | Same as previous Model | ACETON w/wo attention |
| Operationalization | ractors and treatments | Wiodei | ACETON W/ Wo attention |
| | Response variable, elaboration and metric | Partially | Precision and recall per category. Only |
| | nespense vanazie, elaboration and means | . arean, | names |
| | | | Performance: Time (training and |
| | | | prediction) and F1 per time |
| Design | Design type | No | prediction, and 12 per time |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | 10 | 10-fold-cross validation |
| Instrumentation | Test set | Yes | Same as training dataset |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Pytorch. Laptop 2.2.GHz intel core i7 and |
| | | | 16GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Automated constrution of energy test

ID AP37 Experiments 4

Comments E1 hidden (grid search)

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as initial | Does not mention 10-fold cross- validation |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | ACETON | 1 treatment assessment |
| | Response variable, elaboration and metric | Partially | Recall per (missing) category. Only name |
| Design | Design type | No | |
| · · | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training dataset |
| | Measuring instruments | Deduced | ŭ |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | Partially | Pytorch. Laptop 2.2.GHz intel core i7 and 16GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| • | Inferential statistics | No | |
| Validity evaluation | | No | |
| | | | |
| Artifact | Availability | Yes | |

Paper Automated constrution of energy test

ID AP37 Experiments 4

Comments E1 hidden (grid search)

| Aspect | Element | E4 | Comments |
|---------------------|---|--|--|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as initial | Split: 90-100 |
| Operationalization | Training data Factors and treatments | Same as previous Mobile device (how+sw) | Nexus 5X (Android 7.0) vs Nexus 6P |
| Operationalization | Response variable, elaboration and metric | Partially | (Android 6.01.1) Precision, recall |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | Coores 1 min |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training dataset |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced Partially | Dutarch Lantan 2.2 CUs into com 17 |
| | Technological infrastructure | Partially | Pytorch. Laptop 2.2.GHz intel core i7 and 16GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | No | |
| Artifact | A i la la ilita . | Yes | |
| | Availability Badge | No | |

Paper Object detection for graphical user interface: old fashioned or deep learning or a combination

ID AP38 Experiments 7 Comments

| Aspect | Element | E1 | Comments |
|---|--|--|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | No | |
| | Model parameters | No | |
| | DL algorithm | representation: No | |
| | _ | model type: No | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 8-1-1 | 5-fold cross-validation |
| | Training my perparameters | learning rate: No | 5-iola cioss-validation |
| | | #iterations: 160 | |
| | | | |
| | | batch size: 8 | |
| | | #epochs: No | |
| | Training data | Partially | Builds on RICO dataset, but we do not know |
| | | | exactly how |
| Operationalization | Factors and treatments | Hyperparameters | |
| | Response variable, elaboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | | | |
| | Randomization | No | |
| | Task duration | No | |
| | Task duration Procedure | No No | |
| | Task duration Procedure Number of experimental units | No No No | |
| Instrumentation | Task duration Procedure Number of experimental units Test set | No No No Partially | Train set |
| Instrumentation | Task duration Procedure Number of experimental units Test set Measuring instruments | No No No Partially | Train set |
| Instrumentation | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No Partially No | Train set |
| | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No Partially No No No | Train set |
| Population | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No Partially No No No No | Train set |
| | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No Partially No No No No No | Train set |
| Population Analysis | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No Partially No No No No No No No | Train set |
| Population Analysis Validity evaluation | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics Conclusion, internal, construct, external | No No Partially No | Train set |
| Population Analysis | Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No Partially No No No No No No No | Train set |

Paper Object detection for graphical user int

ID AP38 Experiments 7

| _ | | |
|-----|----|------|
| Com | me | ents |

| Aspect | Element | E2 | Comments |
|---|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | · | · | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Approaches. Factor but not | REMAUI, Xianyu, Faster RCNN, YOLOv3, |
| | | treatments | CenterNet. None of them specified |
| | | | |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Response variable, elaboration and metric Design type | Yes No | Precision, recall, F1 |
| Design | · | | Precision, recall, F1 |
| Design | Design type | No | Precision, recall, F1 |
| Design | Design type Blocking variables | No No | Precision, recall, F1 |
| Design | Design type Blocking variables Held-constant variables | No No | Precision, recall, F1 |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No No No | Precision, recall, F1 |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | No No No No | Precision, recall, F1 |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No No No No No No No So | Precision, recall, F1 5-fold cross-validation |
| · | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No No No No No No | |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No No No No No No No So | 5-fold cross-validation |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No No Po No Partially | 5-fold cross-validation |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No No No Portially Deduced | 5-fold cross-validation |
| nstrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No So Partially Deduced Deduced | 5-fold cross-validation |
| nstrumentation Population | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No No So Partially Deduced Deduced No | 5-fold cross-validation |
| nstrumentation Population | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No No No So Partially Deduced Deduced No No | 5-fold cross-validation Train set |
| nstrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics Conclusion, internal, construct, external | No No No No No No So Partially Deduced Deduced No No Partially | 5-fold cross-validation Train set |
| nstrumentation Population Analysis Validity evaluation Artifact | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No No No S Partially Deduced Deduced No No Partially No | 5-fold cross-validation Train set |

Paper Object detection for graphical user int

ID AP38 Experiments 7

| Aspect | Element | E3 | Comments |
|--|---|--|---|
| xperiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Seems 2 factors, nested design | Method (Faster RCNN YOLOv3) and anchor-box settings |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | riocedule | | |
| | Number of experimental units | 5 | 5-fold cross-validation |
| nstrumentation | | | 5-fold cross-validation Train set |
| nstrumentation | Number of experimental units | 5 | |
| nstrumentation | Number of experimental units Test set | 5 Partially | |
| nstrumentation | Number of experimental units Test set Measuring instruments | 5 Partially Deduced | |
| | Number of experimental units Test set Measuring instruments Measurement procedure | 5 Partially Deduced Deduced | |
| Population | Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | 5 Partially Deduced Deduced No | |
| Population | Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | 5 Partially Deduced Deduced No | Train set |
| Population Analysis | Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | 5 Partially Deduced Deduced No No Partially | Train set |
| Instrumentation Population Analysis Validity evaluation Artifact | Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | 5 Partially Deduced Deduced No No Partially No | Train set |

Object detection for graphical user int AP38 Paper

ID 7 Experiments

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| | Trailing data | Same as previous | |
| Operationalization | Factors and treatments | Seems 2 factors | Method (Faster-RCNN, TOLOv3, CenterNet) and amount of training data (2K, 10K, 40K) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | 5 | 5-fold cross-validation |
| Instrumentation | Test set | Partially | Train set |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | 5-fold cross-validation |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Object detection for graphical user int AP38 Paper

ID 7 Experiments

| Aspect | Element | E5 | Comments |
|---------------------|---|------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Seems 2 factors | Method (Faster-RCNN, |
| | | 5005 2 .000.5 | YOLOv3,CenterNet, Our method), Element (nontext-only, mix nontext, text both) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | i recision, recall, i i |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | 5 | 5-fold cross-validation |
| Instrumentation | Test set | Partially | Train set |
| | Measuring instruments | Deduced | |
| | Measurement procedure | Deduced | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | Partially | 5-fold cross-validation |
| • | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | No | |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Object detection for graphical user int

ID AP38 Experiments 7

| Aspect | Element | E5 | Comments |
|---|--|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| | • | · | |
| Operationalization | Factors and treatments | 1 factor | Method (Tesseract, EASET, REMAUI, Xinayu) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | | No | |
| | Blocking variables | No | |
| | Blocking variables Held-constant variables | No | |
| | • | | |
| | Held-constant variables Measured variables (covariates) Randomization | No | |
| | Held-constant variables Measured variables (covariates) | No No | |
| | Held-constant variables Measured variables (covariates) Randomization | No No No No | |
| | Held-constant variables Measured variables (covariates) Randomization Task duration | No No No | 5-fold cross-validation |
| nstrumentation | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No No No No | 5-fold cross-validation Train set |
| nstrumentation | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No No No No Portially Deduced | |
| nstrumentation | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No S Partially | |
| nstrumentation | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No Portially Deduced | |
| | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No No So Partially Deduced Deduced | |
| Population | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No S Partially Deduced Deduced No | |
| Population | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No No No S Partially Deduced Deduced No | Train set |
| Population Analysis | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No S Partially Deduced Deduced No No Partially | Train set |
| nstrumentation Population Analysis Validity evaluation Artifact | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No S Partially Deduced Deduced No No Partially No | Train set |

Paper Object detection for graphical user int

ID AP38 Experiments 7

| Cor | nm | en | ts |
|-----|----|----|----|

| Aspect | Element | E6 | Comments |
|---|---|--|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | No | |
| | Model parameters | No | |
| | DL algorithm | No | |
| | · | | |
| | | | |
| | Training hyperparameters | No | |
| | | | |
| | | | |
| | Training data | Partially | 90,000 GUI elements randomly selected |
| | | | from dataset |
| Operationalization | Factors and treatments | 1 factor | Method (FasterRCNN, YOLOv3, |
| o perationalization | | | |
| o perationalization | | | Centernet, our method) |
| ореганопальная. | | | Centernet, our method) |
| o peranona il autori | | | Centernet, our method) |
| | Response variable, elaboration and metric | yes | Centernet, our method) #bbbox, accuracy, precision, recall, F1 |
| Design | | yes No | |
| | Response variable, elaboration and metric | · | |
| | Response variable, elaboration and metric Design type | No | |
| | Response variable, elaboration and metric Design type Blocking variables | No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables | No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) | No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization | No No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | No No No No No | |
| | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No No No No No No | |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No | #bbbox, accuracy, precision, recall, F1 |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No No No No No No No Partially | #bbbox, accuracy, precision, recall, F1 |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No No No No No No No No Partially Deduced | #bbbox, accuracy, precision, recall, F1 |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No No No No Partially Deduced Deduced | #bbbox, accuracy, precision, recall, F1 |
| Design | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No No No No No Partially Deduced Deduced No | #bbbox, accuracy, precision, recall, F1 |
| Design Instrumentation Population | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No No No No No No No Partially Deduced Deduced No No | #bbbox, accuracy, precision, recall, F1 |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | No No No No No No No No No Partially Deduced Deduced No No No | #bbbox, accuracy, precision, recall, F1 |
| Design Instrumentation Population Analysis | Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No No No No No No Partially Deduced Deduced No No No No | #bbbox, accuracy, precision, recall, F1 |

Paper STATEFORMER: Fine-grained type recovery from binaries using generative state modeling

ID AP39 Experiments 8

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|---|
| Experiment type | | Evaluation | Compares against nothing |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Some | Architecture is described at a high level. Fo |
| | | #neurons/layer: No | details points to supplementary material |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | Does not explicitly mention parameters |
| | • | weights: No | . , . |
| | DL algorithm | representation: Yes | Loss: MSE + BCE. |
| | 224,601.4111 | model type: Yes | Points to supplementary material |
| | | loss function: Yes | romo to supplementary material |
| | | regularization: No | |
| | | • | |
| | Training hunormarameters | optimization: No | Pretrain+train |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | | learning rate: No | Points to supplementary material |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: 10, 50 | |
| | Training data | Yes | Details in supplementary material |
| Operationalization | Factors and treatments | Model type (STATEFORMER) | STATEFORMER performance is evaluated |
| | | | |
| | Response variable, elaboration and metric | Yes | Precision, Recall , F1 |
| Design | Design type | No | |
| | Blocking variables | Deduced | Architecture/optimization/obfuscation |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Details in supplementary material |
| | Measuring instruments | No | • • • • |
| | Measurement procedure | No | |
| | Technological infrastructure | Yes | Pytorch 1.6.0 (Fairseg toolkit) |
| | <u> </u> | | Linux server w Ubuntu 18.04 |
| | | | Intel Xeon 4212 2.2.0GHz 48 virtual cores |
| | | | 188GB RAM |
| | | | 4 Nvidia RTX 2080-Ti GPUs |
| | | | pyelftools, Ghidra |
| Population | Objects (chars. of the experimental datasets) | Yes | Details in supplementary material |
| Analysis | Descriptive statistics | No | Details in supplementary material |
| Allalysis | Inferential statistics | | |
| Validity ovaluation | | No Partially | Flat list of 3 threats |
| Validity evaluation | | Partially | riat iist 01 3 trifeats |
| Artifact | Availability Badge | Yes Yes | Available, reusable |
| | Danse | 1.02 | Available, leusable |

ID AP39 Experiments 8

Comments

Analysis

Artifact

Descriptive statistics

Inferential statistics

Validity evaluation Conclusion, internal, construct, external

Availability

Badge

| Aspect | Element | E2 | Comments |
|---------------------|---|--|--|
| Experiment type | | Evaluation | Against SOTA |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Same as previous??? Model type (STATEFORMER) | Not clear For EKLAVIA, numbers reported in paper are |
| | | v | used. |
| Docien | Response variable, elaboration and metric | Yes No | Accuracy |
| Design | Design type Blocking variables | Deduced | Architecture/optimization |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | Yes | Same 8 projects as EKLAVIA |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Yes | EKLAVIA projects. Supplementary. Material |

Partially

Partially

No

Yes

Yes

Average

Flat list of 3 threats

Available, reusable

AP39 8

| ΕX | per | ım | ents | • |
|----|-----|-----|------|---|
| Co | mn | ner | ıts | |

| Aspect | Element | E3 | Comments |
|---------------------|---|---------------------------------|---|
| Experiment type | | Evaluation | Against SOTA |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous??? | Not clear |
| Operationalization | Factors and treatments | Model type (STATEFORMER, Debin) | Debin already trained model is used. STATEFORMER is restricted to only 17 types, as Debin |
| | Response variable, elaboration and metric | Yes | F1 |
| Design | Design type | No | |
| | Blocking variables | Deduced | Architecture/optimization |
| | | | |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | OpenSSL |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available, reusable |

ID AP39 Experiments 8

| Aspect | Element | E4 | Comments |
|---------------------------------|---|-------------------------------------|--|
| Experiment type | | Evaluation | Against SOTA |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous??? | Not clear |
| Operationalization | Factors and treatments | Model type (STATEFORMER, Typeminer) | Typeminer is not open-source. Authors are contacted and asked for the numbers. It is not DNN |
| | Response variable, elaboration and metric | Yes | F1 |
| Design | Design type Blocking variables | No Deduced (Task) | 1 architecture 1 optimization. The ones used by Typeminer 4 Tasks |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | · | | |
| | | • | menden evaluated en their projects |
| | • | | |
| | Technological infrastructure | Same as previous | |
| Instrumentation | Test set Measuring instruments Measurement procedure Technological infrastructure | Partially No No Same as previous | Mention evaluated on "their" p |
| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | No | |
| | | | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Partially Yes | Flat list of 3 threats |

ID AP39 Experiments 8

| Aspect | Element | E5 | Comments |
|---------------------|---|---------------------------------------|-------------------------------------|
| Experiment type | | Evaluation | Against SOTA |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Madel parameters | Same as previous | |
| | Model parameters | same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | Training data | Same as previous??? | Not clear |
| | Factors and treatments | · · · · · · · · · · · · · · · · · · · | Ghidra is commercial tool (not DNN) |
| | Response variable, elaboration and metric | Yes | Execution time (seconds) |
| Design | Design type | No | |
| | Blocking variables | Project | 4 projects |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Only name the projects |
| | Measuring instruments | No | , and a second |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | | | |
| Depulation | Objects /shows of the sure street day. | Voc | Cumplementary Material |
| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
| Analysis | Descriptive statistics Inferential statistics | Partially No | Average |
| Validity ovaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | Yes | riat iist UI 3 tiileats |
| Aidiact | Badge | Yes | Available, reusable |
| | Dauge | 163 | Available, leasable |

ID AP39 Experiments 8

| Aspect | Element | _E6 | Comments |
|---------------------|---|------------------------------|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous??? | Not clear |
| Operationalization | Factors and treatments | Use of pre-training, masking | Not sure the value of the other once one of them is fixed |
| | Response variable, elaboration and metric | Yes | F1 |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | Technological infrastructure | Same as previous | |
| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available, reusable |

ID AP39 Experiments 8

| Aspect | Element | E7 | Comments |
|---------------------|---|---------------------------|-------------------------|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | Training data | Same as previous??? | Not clear |
| Operationalization | Factors and treatments | Assesses STATEFORMER only | |
| | Response variable, elaboration and metric | Yes (pre-training loss) | MSE, BCE |
| Design | Design type | No | WISE, BEE |
| Design | Blocking variables | No | |
| | Sissimily remarks | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | Yes | |
| | | | |

ID AP39 Experiments 8

| Aspect | Element | E8 | Comments |
|---------------------|---|---|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Same as previous??? Pre-training (no, STATEFORMER, TREX) | Not clear TREX is DNN, but they do not mention where they take it from |
| | Response variable, elaboration and metric | Yes | F1 |
| Design | Design type | No | 1.1 |
| 8 | Blocking variables | No | |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| nstrumentation | Test set | No | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Same as previous | |

| Population | Objects (chars. of the experimental datasets) | Yes | Supplementary. Material |
|---------------------|---|-----------|-------------------------|
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list of 3 threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available, reusable |

Paper A syntax-guided edit decoder for neural program repair

ID AP40 Comments

5 Experiments

| Aspect | Element | E1 | Comments |
|---------------------|---|-----------------------------|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Not clear | |
| | | #neurons/layer: | |
| | | connections: No | |
| | | activation functions: Some | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | Loss function: maximize negative log- |
| | | model type: Yes | likelihood of the oracle edit sequence |
| | | loss function: Yes | |
| | | regularization: Dropout 0.1 | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 80-20 | |
| | | learning rate: 0.0001 | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Partially | Explanation. Could be reproduced, but it is |
| | | | not explicitly linked |
| Operationalization | Factors and treatments | Approaches. Factors but not | jGenProg, HDRepair, Nopol, CapGen, |
| | | treatments | SketchFix, FixMiner, SimFix, Tbar, DLFix, |
| | | | PraPR, AVATAR, Recoder |
| | Response variable, elaboration and metric | Yes | Number of correct patches without perfect |
| | | | fault localization |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | 5 hours | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Defects4J v1.2. Described but not explicitly |
| | | | linked to artifact |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | No characteristics are provided |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Only external and internal |
| Artifact | Availability | Yes | |
| | Badge | Available | |

 Paper
 A syntax-guided edit decoder for neural prograr

ID AP40 Experiments 5

| Aspect | Element | E2 | Comments |
|---------------------|---|---|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | same as previous | |
| | . 0 | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| Operationalization | Factors and treatments | SequenceR, CODIT, DLFix, CoCoNuT, TBar, Recoder | |
| | | | |
| | Response variable, elaboration and metric | Number of correct patches with perfect fault localization | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | 5 hours | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Defects4J v1.2. Described but not explicitly |
| | | , | linked to artifact |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | No characteristics are provided |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Only external and internal |
| Artifact | Availability | Yes | |
| | Badge | Available | |

Paper A syntax-guided edit decoder for neural prograr

Badge

ID AP40 5 Experiments

Comments

| Aspect | Element | E3 | Comments |
|---------------------|--|---|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | · | |
| | | | |
| | Model parameters | Same as previous | |
| | | | |
| | DL algorithm | same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | · | |
| | Training data | Same as previous | |
| | | | |
| Operationalization | Factors and treatments | Removing: modify, subtreecopy, insert, placeholder. With eveverything | But testsets are not expected in the code It cannot be FA |
| | Response variable, elaboration and metric | Number of correct patches without perfect fault localization | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | 5 hours | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Defects4J v1.2. Described but not explicitly linked to artifact |
| | Measuring instruments | No | explicitly liliked to artifact |
| | | | |
| | Measurement procedure Technological infrastructure | No No | |
| Population | Objects (chars. of the experimental datasets) | No | No characteristics are provided |
| • | | No | no characteristics are provided |
| Analysis | Descriptive statistics | No | |
| Validity ovalvation | Inferential statistics | | Only outernal and internal |
| Artifact | Conclusion, internal, construct, external | Partially | Only external and internal |
| ArtiidCl | Availability | Yes | |

Available

Paper A syntax-guided edit decoder for neural prograr

ID AP40 Experiments 5

| Aspect | Element | E4 | Comments |
|---------------------|---|--|--|
| Experiment type | | Generalization/Evaluation | Tested in a diferent dataset |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | Training data | Camp or proving | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Tbar, SimFix, Decoder | |
| | Response variable, elaboration and metric | Number of correct patches without perfect fault localization | |
| Design | Design type | No | |
| _ | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | 5 hours | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Partially | Defects4J 2.0. Described but not explicitly linked to artifact |
| | Measuring instruments | No | . , |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Dalatia | Objects (chars. of the experimental datasets) | No | No characteristics are provided |
| Population | = =,=== (onaior or are experimental datasets) | | sharacteriotics are provided |
| | Descriptive statistics | No | |
| | Descriptive statistics Inferential statistics | No No | |
| Analysis | Inferential statistics | No | Only external and internal |
| Analysis | · · · · · · · · · · · · · · · · · · · | | Only external and internal |

 Paper
 A syntax-guided edit decoder for neural prograr

ID AP40 Experiments 5

| Aspect | Element | E5 | Comments |
|---------------------|---|---|---|
| Experiment type | | Optimization | Diferent sizes of training dataset |
| Hypotheses | Research hypotheses | No | No associated RQ in paper |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Different sizes of training set: 25%, 50%, 75%, 85%, 90%, 93% 96%, 100% | But testsets are not expected in the code Therefore, we do not know which partitions exactly have been chosen |
| | Response variable, elaboration and metric | Number of correct patches without perfect fault localization | , |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | 5 hours | |
| | Procedure | No | |
| | Number of experimental units | Yes | 5 runs |
| Instrumentation | Test set | Partially | Defects4J v1.2. Described but not explicitly linked to artifact |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | No characteristics are provided |
| Analysis | Descriptive statistics | Yes | Boxplot |
| . , | Inferential statistics | No | - 1 |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Only external and internal |
| Artifact | Availability | Yes | z, external and meeting |
| | · | | |
| | Badge | Available | |

Paper

 $Light weight global \ and \ local \ contexts \ guided \ method \ name \ recommendation \ with \ prior \ knowledge$

ID Experiments AP41 7

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------|--------------------------------------|
| Experiment type | | Optimization | Hidden |
| I borothooo | December 1991 | Nie | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as next | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as next | |
| | Woder parameters | Same as next | |
| | DL algorithm | Same as next | |
| | DE digonalin | Same as nexe | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as next | |
| | g, perparameters | ourine do nexe | |
| | | | |
| | | | |
| | | | |
| | Training data | No | |
| | | | |
| | | | |
| Operationalization | Factors and treatments | Yes | Number of tokens from implementation |
| | | | context (5,10,20) |
| | | | |
| | Response variable, elaboration and metric | No | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | No | |
| | | | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | |
| Aitilact | Badge | Yes | Available |

ID AP41 Experiments 7

| - | | | |
|-----|---|----|-----|
| _ | | | |
| Com | m | er | 1ts |

| Aspect | Element | E2 | Comments |
|---------------------|---|------------------------------|--|
| Experiment type | | Evaluation | Compares against SOTA |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | |
| | | #neurons/layer: No | |
| | | connections: No | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | "Due to page limit, we only briefly introduce this |
| | _ | model type: Yes | model in the paper, and more details could be |
| | | loss function: Yes | referred to the existing work [57]" |
| | | regularization: No | , , , , , , , , , , , , , , , , , , , |
| | | optimization: No | Loss function:negative log likelihood of the |
| | | optimization. No | oracle word for that step |
| | Training hyperparameters | train-test split: Yes | oracle word for triat step |
| | rialining hyperparameters | learning rate: No | |
| | | #iterations: No | |
| | | | |
| | | batch size: No | |
| | | #epochs: No | 5.6 |
| | Training data | Yes | References known datasets (Java-small, Java- |
| | | | med, Java-large, Mnire) publicly available |
| Operationalization | Factors and treatments | Model type (10 approaches vs | For Mnire, they use results reported in paper. Do |
| | | Cognac) | not mention other approaches (could be the |
| | | | same) |
| | Response variable, elaboration and metric | Yes | Precision, Recall, F-score with formulas |
| Design | Design type | No | |
| | Blocking variables | Dataset | Report results per dataset. Could be factor? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training sets. They train and test with |
| | | | the same test. |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| | Descriptive statistics | No | |
| Analysis | · | | Conclusions based on 1 run. At a guess |
| Validity and best | Inferential statistics | No Portiol | Conclusions based on 1 run. At a guess |
| • | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | A - 21.11 |
| | Badge | Yes | Available |

ID AP41 Experiments 7

| Aspect | Element | E3 | Comments |
|---------------------|---|---|--|
| Experiment type | | Generalization/Evaluation | Compares agains SOTA for other task |
| | | | (inconsistencies detection |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Yes | Reference to known dataset (Liu et al), publicly available |
| Operationalization | Factors and treatments | Model type (Liu et al, Mnire, Cognac) Class (consistent, inconsistent) | Do not explain any of the others |
| | Response variable, elaboration and metric | Yes | Precision, Recall, F-score |
| Design | Design type | No | |
| 200,6.1 | Blocking variables | No | Could class be a blocking variable? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Conclusions based on 1 run. At a guess |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |
| | | | |

ID AP41 Experiments 7

| Aspect | Element | E4 | Comments |
|---------------------|---|--|--|
| Experiment type | | Optimization | Ablation study for the task in E2 |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Yes | Same as E2 |
| Operationalization | Factors and treatments | Model (no caller info, no callee info, no prior knowledge, Cognac) | No further details are given |
| | Response variable, elaboration and metric | Yes | F-score |
| Design | Design type | No | |
| | Blocking variables | Dataset | Report results per dataset. Could be factor? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training sets. They train and test with the same test. |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Conclusions based on 1 run. At a guess |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |
| | | | |

Badge

ID AP41 Experiments 7

Comments

| Aspect | Element | E5 | Comments |
|---------------------|---|--|--|
| Experiment type | | Optimization | Ablation study for the task in E3 |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Yes | Same as E3 |
| Operationalization | Factors and treatments | Model (no caller info, no callee info, no prior knowledge, Cognac) Class (consistent/inconsistent) | No further details are given |
| | Response variable, elaboration and metric | Yes | F-score, Accuracy |
| Design | Design type | No | |
| | Blocking variables | No | Could class be a blocking variable? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Conclusions based on 1 run. At a guess |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | |

Yes

Available

ID AP41 Experiments 7

| Aspect | Element | E6 | Comments |
|---------------------------------|--|-------------------------------------|---|
| Experiment type | | Optimization | use of caller/calle info. |
| Hypotheses | Research hypotheses | No | |
| riypotrieses | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| variables selection | wiodernyperparameters | Salite as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Yes | References known dataset (Mnire), publicly available |
| Operationalization | Factors and treatments | Model type (seq2seq model vs. Cogna | No details are given about seq2seq |
| | Response variable, elaboration and metric | Yes | F-score |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training set. |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Conclusions based on 1 run. At a guess |
| | | | |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Partial Yes | Flat list of threats |

ID AP41 Experiments 7

| • | |
|--------|----|
| Commen | +- |

| Aspect | Element | E7 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Optimization | Similar to E1, but with all tokens |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Yes | References known datasets(Java-small, Java-med, Java-large, Mnire), publicly available |
| Operationalization | Factors and treatments | Yes | Tokens (all vs. 10) |
| | Response variable, elaboration and metric | Yes | F-score |
| Design | Design type | No | |
| | Blocking variables | Dataset | Report results per dataset. Could be factor? |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training sets. They train and test with the same test. |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | No | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Conclusions based on 1 run. At a guess |
| Validity evaluation | Conclusion, internal, construct, external | Partial | Flat list of threats |
| Artifact | Availability | Yes | |
| | Badge | Yes | Available |

Paper Automating the removal of obsolete TODO comments

ID AP42 Experiments 3

Comments Wild study is a case study, not a controlled experiment

| Aspect | Element | E1 | Comments |
|---------------------|---|------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 12x3 (BERT)+ 3 + 1 | Plus 3 encoders (TODO comment, code |
| | | #neurons/layer: embeddings: | change, commit message) that generate |
| | | 768x3, rest ??? | embeddings using pre-trained BERT. The 3 |
| | | connections: Yes | encoders are jointly trained with the DNN. |
| | | activation functions: ReLu | |
| | | params. Initialization: | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: cross entropy | |
| | | regularization: dropout 0.2 | |
| | | dense layers MLP, clip the | |
| | | gradients norm by 2 | |
| | | optimization: Adam | |
| | Training hyperparameters | train-test split: 80-10-10 | |
| | Training Try perparameters | learning rate: 0.001 | |
| | | #iterations: 1,000 | |
| | | batch size: 32 | |
| | | | |
| | Training data | #epochs: No Yes | Included in RP |
| Operationalization | Training data Factors and treatments | Partially | Compares against SOTA (TCO, TMO, TCMO, |
| Operationalization | ractors and deadnerits | ratually | IRSC, TDCleaner) |
| | | | inde, i beleaner) |
| | Response variable, elaboration and metric | Yes | Accuracy, precision, recall, F1, with formulas |
| | nesponse variable, classification and metric | 163 | Accuracy, precision, recall, 1-1, with formulas |
| Design | Design type | No | |
| | Blocking variables | Deduced | Dataset (Python, Java) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Included in RP |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Hardware missing. Python, using Pytorch |
| | | | (versions?) |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in paper |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Based on single values |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Automating the removal of obsolete T

ID AP42 Experiments 3

Comments Wild study is a case study, not a contr

| Aspect | Element | E2 | Comments |
|-------------------------------------|---|--|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | Divide a Object | C | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Training data | Yes | Included in RP |
| Operationalization | Factors and treatments | Partially | Removes encoders (without commit mesage, without code change, without comment, TDCleaner) |
| | Response variable, elaboration and metric | Yes | Accuracy, precision, recall, F1, with formulas |
| Design | Design type | No | |
| | Blocking variables | Deduced | Dataset (Python, Java) |
| | | | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No No | |
| | Measured variables (covariates) Randomization | | |
| | Measured variables (covariates) Randomization Task duration | No No | |
| | Measured variables (covariates) Randomization Task duration Procedure | No No No | |
| | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No No No No | Seems 1 run |
| nstrumentation | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No No No No Yes | Seems 1 run Included in RP |
| Instrumentation | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No No No No Yes | |
| Instrumentation | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No Yes No No | Included in RP |
| Instrumentation | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No No No No Yes | |
| | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No No No No Yes No No | Included in RP Hardware missing. Python, using Pytorch |
| Instrumentation Population Analysis | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No No No No Yes No No Partially | Included in RP Hardware missing. Python, using Pytorch (versions?) |
| Population | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No No No No Yes No Partially | Included in RP Hardware missing. Python, using Pytorch (versions?) |
| Population | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No Yes No Partially Yes No | Included in RP Hardware missing. Python, using Pytorch (versions?) Described in paper |
| Population Analysis | Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No No No No Yes No No Partially Yes No No | Included in RP Hardware missing. Python, using Pytorch (versions?) Described in paper Based on single values |

Paper Automating the removal of obsolete T

ID AP42 Experiments 3

Comments Wild study is a case study, not a contr

| Experiment type Optimization Replaces BERT with Word2Vec Hypotheses Yes RQ3 Variables selection Model hyperparameters Same as previous Variables selection Model parameters Same as previous Model parameters Same as previous Training hyperparameters Same as previous Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Pesign Pesign type Accuracy, precision, recall, F1, with formulas Design Design type No Accuracy, precision, recall, F1, with formulas Measured variables (elaboration and metric left-constant variables (elaboration and metric left-constant variables) Yes Accuracy, precision, recall, F1, with formulas Design Design type No Deduced Dataset (Python, Java) Measured variables (covariates) No Design (elaboration and metric left-constant variables) No Measured variables (covariates) No Design (elaboration and metric left-constant variables) No Procedure No No | Aspect | Element | E3 | Comments |
|--|---------------------|---|------------------|-----------------------------|
| Variables selection Model hyperparameters Same as previous | Experiment type | | Optimization | Replaces BERT with Word2Vec |
| Variables selection Model hyperparameters | Hypotheses | Research hypotheses | Yes | RQ3 |
| Model parameters Same as previous DL algorithm Same as previous Training hyperparameters Same as previous Training data Yes Included in RP Partially Replaces BERT with WordZVec Response variable, elaboration and metric Response variable, elaboration and metric Response variable, elaboration and metric Pesign Design VPE Blocking variables Blocking variables Peduced Blocking variables Measured variables (covariates) Measured variables (covariates) No Randomization Task duration No Task duration No Procedure No No Task duration No Procedure No No Reasured variables No Reasured variables (covariates) No Reasured variables (covariates) No Reasured variables (covariates) No Reasured variables (covariates) No Reasured variables (reprinciple (reprinc | | Statistical hypotheses | No | |
| DL algorithm Same as previous Training hyperparameters Same as previous Training data Yes Included in RP Training data Yes Accuracy, precision, recall, F1, with formulas Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Response variables No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Mandomization No Task duration No Procedue No Task duration No Procedue No No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Population Objects (chars. of the experimental datasets) No Based on single values Artifact Availability Partially Fiatlist | Variables selection | Model hyperparameters | Same as previous | |
| DL algorithm Same as previous Training hyperparameters Same as previous Training data Yes Included in RP Training data Yes Accuracy, precision, recall, F1, with formulas Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Response variables No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Mandomization No Task duration No Procedue No Task duration No Procedue No No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Population Objects (chars. of the experimental datasets) No Based on single values Artifact Availability Partially Fiatlist | | | | |
| DL algorithm Same as previous Training hyperparameters Same as previous Training data Yes Included in RP Training data Yes Accuracy, precision, recall, F1, with formulas Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Response variables No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Mandomization No Task duration No Procedue No Task duration No Procedue No No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Population Objects (chars. of the experimental datasets) No Based on single values Artifact Availability Partially Fiatlist | | | | |
| Training hyperparameters Same as previous Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Pessign Design Design type Blocking variables Held-constant variables Held-constant variables Held-constant variables No Measured variables(covariates) Randomization No Task duration Procedure No No No Task duration No Task duration Procedure No No No No Heasuring instruments Measurement procedure No Measurement No Measurement No Measurement | | Model parameters | Same as previous | |
| Training hyperparameters Same as previous Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Pessign Design Design type Blocking variables Held-constant variables Held-constant variables Held-constant variables No Measured variables(covariates) Randomization No Task duration Procedure No No No Task duration No Task duration Procedure No No No No Heasuring instruments Measurement procedure No Measurement No Measurement No Measurement | | DL algorithm | Same as previous | |
| Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Randomization No Procedure No No Member of experimental units No Seems 1 run Instrumentation Test et Measuring instruments No Measurement procedure No Measurement N | | | | |
| Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Randomization No Procedure No No Member of experimental units No Seems 1 run Instrumentation Test et Measuring instruments No Measurement procedure No Measurement N | | | | |
| Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Randomization No Procedure No No Member of experimental units No Seems 1 run Instrumentation Test et Measuring instruments No Measurement procedure No Measurement N | | | | |
| Training data Yes Included in RP Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Randomization No Procedure No No Member of experimental units No Seems 1 run Instrumentation Test et Measuring instruments No Measurement procedure No Measurement N | | Training hypernarameters | Same as previous | |
| Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Blocking variables Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Measuring instruments No Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Based on single values Validity evaluation Conclusion, internal, construct, external Partially Flat list Artifact Availability Yes | | Training Hyperparameters | Same as previous | |
| Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Blocking variables Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Measuring instruments No Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Based on single values Validity evaluation Conclusion, internal, construct, external Partially Flat list Artifact Availability Yes | | | | |
| Operationalization Factors and treatments Partially Replaces BERT with Word2Vec Response variable, elaboration and metric Yes Accuracy, precision, recall, F1, with formulas Design Design type No Blocking variables Deduced Dataset (Python, Java) Held-constant variables No Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Measuring instruments No Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Based on single values Validity evaluation Conclusion, internal, construct, external Partially Flat list Artifact Availability Yes | | Training data | Yes | Included in RP |
| Response variable, elaboration and metric Pesign Design Design type Blocking variables Held-constant variables Measured variables Measured variables Randomization Randomization No Randomization No Randomization No Procedure Number of experimental units No | Operationalization | | | |
| Design typeNoDesign typeDeducedDataset (Python, Java)Held-constant variablesNoMeasured variables (covariates)NoRandomizationNoTask durationNoProcedureNoNumber of experimental unitsNoMeasuring instrumentsNoMeasuring instrumentsNoMeasurement procedureNoMeasurement procedureNoMeasurement procedureNoTechnological infrastructurePartiallyHardware missing. Python, using Pytorch (versions?)PopulationObjects (chars. of the experimental datasets)YesDescribed in paperAnalysisDescriptive statisticsNoValidity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | Operationalization | ractors and treatments | raruany | Replaces BERT With Word2Vec |
| Design typeNoDesign typeDeducedDataset (Python, Java)Held-constant variablesNoMeasured variables (covariates)NoRandomizationNoTask durationNoProcedureNoNumber of experimental unitsNoMeasuring instrumentsNoMeasuring instrumentsNoMeasurement procedureNoMeasurement procedureNoMeasurement procedureNoTechnological infrastructurePartiallyHardware missing. Python, using Pytorch (versions?)PopulationObjects (chars. of the experimental datasets)YesDescribed in paperAnalysisDescriptive statisticsNoValidity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | | | | |
| Design per | | Response variable, elaboration and metric | Yes | * * |
| Blocking variables Deduced Dataset (Python, Java) | | | | formulas |
| Held-constant variables Measured variables (covariates) No Randomization No Task duration No Procedure No Number of experimental units No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure No Measurement procedure No Technological infrastructure Partially Population Objects (chars. of the experimental datasets) No Poscriptive statistics No Inferential statistics No Onclusion, internal, construct, external Artifact Availability Artifact No | Design | | | |
| Measured variables (covariates) Randomization No Task duration No Procedure No Number of experimental units No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure No Technological infrastructure Partially Partially No Population Objects (chars. of the experimental datasets) No Inferential statistics No Inferential statistics No Oseriptive | | _ | | Dataset (Python, Java) |
| Randomization No Task duration No Procedure No No Number of experimental units No Seems 1 run Instrumentation Test set Yes Included in RP Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially Artifact Availability Yes | | | | |
| Task duration No Procedure No Number of experimental units No Seems 1 run Instrumentation Test set Yes Included in RP Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Inferential statistics No Validity evaluation Conclusion, internal, construct, external Partially Artifact Availability Yes | | | | |
| Procedure Number of experimental units No Seems 1 run Instrumentation Test set Measuring instruments Measurement procedure Technological infrastructure Partially Population Objects (chars. of the experimental datasets) No Analysis Descriptive statistics No Inferential statistics No Validity evaluation Availability Availability No Seems 1 run Included in RP Hardware missing. Python, using Pytorch (versions?) Population Described in paper Flat list Flat list Flat list | | | | |
| Number of experimental unitsNoSeems 1 runInstrumentation Instrumentation Measuring instruments Measurement procedure Technological infrastructureNoIncluded in RPPopulationNoHardware missing. Python, using Pytorch (versions?)PopulationObjects (chars. of the experimental datasets)YesDescribed in paperAnalysisDescriptive statistics Inferential statisticsNoBased on single valuesValidity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | | | | |
| Instrumentation Measuring instruments No Measurement procedure No Technological infrastructure Partially Hardware missing. Python, using Pytorch (versions?) Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Based on single values Validity evaluation Conclusion, internal, construct, external Partially Artifact Availability Yes | | | | C 1 |
| Measuring instruments Measurement procedure Technological infrastructure Partially Population Objects (chars. of the experimental datasets) No Poscriptive statistics No Inferential statistics No Validity evaluation Availability Availability No | | • | | |
| Measurement procedure Technological infrastructure Population Objects (chars. of the experimental datasets) Population Objects (chars. of the experimental datasets) Objects (chars. of the experimental datas | Instrumentation | | | Included in RP |
| PopulationObjects (chars. of the experimental datasets)YesDescribed in paperAnalysisDescriptive statisticsNoValidity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | | • | | |
| Population Objects (chars. of the experimental datasets) Yes Described in paper Analysis Descriptive statistics No Inferential statistics No Based on single values Validity evaluation Conclusion, internal, construct, external Partially Flat list Artifact Availability Yes | | | | |
| AnalysisDescriptive statisticsNoInferential statisticsNoBased on single valuesValidity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | | Technological infrastructure | Partially | |
| Validity evaluationConclusion, internal, construct, externalPartiallyFlat listArtifactAvailabilityYes | Population | Objects (chars. of the experimental datasets) | Yes | Described in paper |
| Validity evaluation Conclusion, internal, construct, external Partially Flat list Artifact Availability Yes | Analysis | Descriptive statistics | No | |
| Artifact Availability Yes | | Inferential statistics | No | Based on single values |
| • | Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Badge No | Artifact | Availability | Yes | |
| | | Badge | No | |

Paper Vulnerability detection with fine-grained interpretations

ID AP43 Experiments 4

Comments R1 cross-validation is a different experiment, RQ2 experiments with a component outside the DNN, RQ3

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ6 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: Yes | |
| | | #neurons/layer: No | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: 80-10-10 | Divide 80-10-10 vulnerable method. For |
| | | learning rate: No | training, the same number of non- |
| | | #iterations: No | vulnerable methods are added. For |
| | | batch size: No | validation/testing, real ratio between |
| | | #epochs: No | vulnerable/not vulnerable used |
| | Training data | Partially | Referenced (datasets used by others) but |
| | Training data | raitially | , , |
| | | | not explicitly linked to artifact (see |
| | | | data/results of experiments). Fan et al, |
| 0 | Factors and the state of the | Desate II. | Reveal, FFMPeg+Qemu |
| Operationalization | Factors and treatments | Partially | Model type (VulDeePecker, Devign, SyseVR, |
| | | | Russel, Reveal, IVDetect). All DNNs |
| | Persons variable elaboration and matric | Yes | Maan average precision normalized DCC |
| | Response variable, elaboration and metric | 165 | Mean average precision, normalized DCG, |
| | | | first ranking, accuracy under curve, |
| | | | precision, recall, F-score, training and |
| Dasian | Design trung | Ne | prediction time for IVDetect |
| Design | Design type Blocking variables | No Deduced | Dataset |
| | | | Dataset |
| | Held-constant variables Measured variables (covariates) | No No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Same as training | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Characteristics explained |
| Analysis | Descriptive statistics | No | · |
| • | Inferential statistics | No | Based on single values |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Just one threat |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Vulnerability detection with fine-grain

ID AP43 Experiments 4

Comments R1 cross-validation is a different expeis not an experiment

| Aspect | Element | E2 | Comments |
|---------------------|---|---|--|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | train-test split: cross-validation learning rate: No | Training: Reveal, FFMPeg+Qemu, 20% Fan, testing: 80% fan |
| | | #iterations: No | |
| | | batch size: No | |
| | Testistics date | #epochs: No | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | training type (within/cross) on IVDetect |
| Operationalization | ractors and treatments | Partially | training type (within) cross) on type tect |
| | Response variable, elaboration and metric | Yes | Mean average precision, normalized DCG |
| Docign | Design type | No | |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Same as training | 80% Fan |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Characteristics explained |
| Analysis | Descriptive statistics | No | |
| - | Inferential statistics | No | Based on single values |
| Validity evaluation | Conclusion, internal, construct, external | Partially | |
| | | Yes | |
| Artifact | Availability | 162 | |

Paper Vulnerability detection with fine-grain

ID AP43 Experiments 4

Comments R1 cross-validation is a different expe

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as E1 | |
| | | | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | Model type (ST, SST, AST, Var, CD, IVDetect) |
| | | • | |
| | Response variable, elaboration and metric | Yes | Mean average precision, normalized DCG, first ranking, accuracy under curve, precision, recall, F-score |
| Design | Design type | No | |
| 0 | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Same as training | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Characteristics explained |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Based on single values |
| Validity evaluation | Conclusion, internal, construct, external | Partially | |
| | Availability | Yes | |
| Artifact | Availability | 163 | |

Paper Vulnerability detection with fine-grain

ID AP43 Experiments 4

Comments R1 cross-validation is a different expe

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|---|
| Experiment type | | Optimization | Factor is train-test-split |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Mandal and a service state of | C | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as E1 | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | Training data | Same as previous | |
| | | | |
| | | | |
| Operationalization | Factors and treatments | Yes | Train/test split (80-10-10, 70-15-15, 60- |
| | | | 20-20, 50-25-25) |
| | | | |
| | Response variable, elaboration and metric | Yes | Mean average precision, normalized |
| | | | DCG, first ranking, accuracy under curve, |
| | | | precision, recall, F-score |
| Design | Design type | No | |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Same as training | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Characteristics explained |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | Based on single values |
| Validity evaluation | Conclusion, internal, construct, external | Partially | |
| | | | |
| Artifact | Availability | Yes | |

Paper Data-driven accessibility repair revisited: on the effectiveness of generating labels for icons in android apps

ID AP44 Experiments 1

Comments RQ4 and RQ5 experiment with LabelDroid (a SOTA proposal)

| Aspect | Element | E1 | Comments |
|-------------------------------------|--|--|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ6, RQ8 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: LSTM (context encoder 4 layers), ResNet (image encoder), fully connected, LSTM (decoder 4 layers), Softmax #neurons/layer: 63 bits input (?) connections: In RP activation functions: In RP params. Initialization: In RP | Pre-trained ResNet18. One hot encoder and GloVe |
| | Model parameters | biases: No weights: No | |
| | DL algorithm | representation: Yes model type: Yes loss function: (weighted?) cross- entropy regularization: In RP optimization: Adam | |
| | Training hyperparameters | train-test split: 80-10-10 learning rate: RP #iterations: In RP batch size: In RP #epochs: In RP | Guided grid search |
| | Training data | Yes | Images extracted from Rico dataset |
| Operationalization | Factors and treatments Response variable, elaboration and metric | Partially Partially | COALA, LabelDroid. In theory, COALA should be (almost) fully defined in RP, but LabelDroid is not BLEU, METEOR, ROUGH, CIDEr, exact match, time (for COALA only). Formulas missing |
| | | | |
| Design | Design type | No | |
| Design | Design type Blocking variables | No Deduced | Random split (5 times) |
| Design | | | Random split (5 times) |
| Design | Blocking variables | Deduced | Random split (5 times) |
| Design | Blocking variables Held-constant variables | Deduced No | Random split (5 times) |
| Design | Blocking variables Held-constant variables Measured variables (covariates) | Deduced No No | Random split (5 times) |
| Design | Blocking variables Held-constant variables Measured variables (covariates) Randomization | Deduced No No No | Random split (5 times) NLGE Python library |
| Design | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | Deduced No No No No | NLGE Python library |
| | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration | Deduced No No No No Yes | |
| | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | Deduced No No No Vo Yes | NLGE Python library Seems 1 run (per random split) Same as training |
| | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | Deduced No No No No Yes No Yes Yes | NLGE Python library Seems 1 run (per random split) |
| Design | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | Deduced No No No No No No Yes | NLGE Python library Seems 1 run (per random split) Same as training |
| | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | Deduced No No No No Yes No Yes Yes Deduced | NLGE Python library Seems 1 run (per random split) Same as training NLGE Python library PyTorch. Ubuntu wiht NVIDIA GP102 GPU |
| Instrumentation | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | Deduced No No No No Yes No Yes Deduced Partially Yes | NLGE Python library Seems 1 run (per random split) Same as training NLGE Python library PyTorch. Ubuntu wiht NVIDIA GP102 GPU and 128GB memory Details shown |
| Instrumentation | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Deduced No No No No Yes No Yes Deduced Partially Yes Partially | NLGE Python library Seems 1 run (per random split) Same as training NLGE Python library PyTorch. Ubuntu wiht NVIDIA GP102 GPU and 128GB memory |
| Instrumentation Population Analysis | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | Deduced No No No No No Yes No Yes Deduced Partially Yes Partially No | NLGE Python library Seems 1 run (per random split) Same as training NLGE Python library PyTorch. Ubuntu wiht NVIDIA GP102 GPU and 128GB memory Details shown Value used is mean(?) due to 5 times |
| Instrumentation | Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics | Deduced No No No No Yes No Yes Deduced Partially Yes Partially | NLGE Python library Seems 1 run (per random split) Same as training NLGE Python library PyTorch. Ubuntu wiht NVIDIA GP102 GPU and 128GB memory Details shown |

Paper Boosting coverage-based fault localization via graph-based representation learning

ID AP45 Experiments 4

| Aspect | Element | E1 | Comments |
|------------------------------|--|---------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 8(?) Embedding, | Mention layer normalization and residual |
| | | GGNN (5 layers?), fully (?), | connection |
| | | softmax | |
| | | #neurons/layer: No | Embedding size: 32 |
| | | connections: Yes | |
| | | activation functions: sigmoid, | |
| | | tanh (others?) | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: listwise ranking | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | |
| | | learning rate: 0.01 | |
| | | #iterations: No | |
| | | batch size: 60 (20 for closure | |
| | | project) | |
| | | #epochs: 10 | |
| | Training data | Yes | Defects4J (V1.2.0). Publicly available |
| Operationalization | Factors and treatments | Partially | Ochiai, CNNFL, FLUCCS (no DNNs) DeepFL, Grace (DNNs) |
| | Response variable, elaboration and metric | Yes | Recall at Top-N, MFR, MAR, time (for Grace) |
| Design | Design type | No | |
| Design | Blocking variables | Deduced | leave-one-out cross validation |
| | Held-constant variables | No | Fixed random seeds |
| | Measured variables (covariates) | No | r incu rundom seeds |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | · · |
| | Measurement procedure | No | |
| | Technological infrastructure | Yes!!!! | Dell, 300 GB RAM, Intel Xeon CPU E5-2680 |
| | | | v4 @2.40 GHz, and 8 24GB GPUs of GEForce |
| | | | RTX3090, running Ubuntu 16.04.6 LTS. |
| | | | PyTorch V1.7.1 |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in Table 2 |
| Analysis | Descriptive statistics | Partially | Averages??? (cross validation) |
| Alidiysis | Inferential statistics | Yes | Wilcoxon |
| • | | | |
| Validity evaluation | | Partially | Missing conclusion |
| Validity evaluation Artifact | Conclusion, internal, construct, external Availability | Partially Yes | Missing conclusion |

Paper Boosting coverage-based fault localiza

Availability

Badge

Artifact

ID AP45 Experiments 4

Comments

| Aspect | Element | E2 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Training data | Yes | Defects4J (V1.2.0). Publicly available |
| Operationalization | Factors and treatments | Yes | loss function (listwise, pairwise, pointwise), code represntation (2 variants), test representation (2 variants) |
| | Response variable, elaboration and metric | Yes | MFR, MAR |
| Design | Design type | No | |
| | Blocking variables | Deduced | leave-one-out cross validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Yes!!!! | Dell, 300 GB RAM, Intel Xeon CPU E5-2680 v4 |
| | | | @2.40 GHz, and 8 24GB GPUs of GEForce |
| | | | RTX3090, running Ubuntu 16.04.6 LTS. PyTorch |
| | | | V1.7.1 |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in Table 2 |
| Analysis | Descriptive statistics Inferential statistics | Partially No | Averages??? (cross validation) |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing conclusion |
| A stifact | Availability | Voc | WIDSHIE COLICIOSION |

Yes

No

Paper Boosting coverage-based fault localiza

ID AP45 Experiments 4

Badge

Comments

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Generalization | Integrated with other techniques |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Yes Partially | Defects4J (V1.2.0). Publicly available DeepFL, DeepFL+Grace |
| | Response variable, elaboration and metric | Yes | Recall at Top-N, MFR, MAR, time (for Grace) |
| Design | Design type | No | |
| | Blocking variables | Deduced | leave-one-out cross validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Yes!!!! | Dell, 300 GB RAM, Intel Xeon CPU E5-2680 v4 @2.40 GHz, and 8 24GB GPUs of GEForce RTX3090, running Ubuntu 16.04.6 LTS. PyTorch V1.7.1 |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in Table 2 |
| Analysis | Descriptive statistics | Partially | Averages??? (cross validation) |
| | Inferential statistics | Yes | Wilcoxon |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing conclusion |
| Artifact | Availability | Yes | |
| | Padra | Ne | |

No

Paper Boosting coverage-based fault localiza

ID AP45 Experiments 4

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Generalization | Cross-project |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | Training Hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Training data | Yes | Defects4J (V1.2.0). Publicly available |
| Operationalization | Factors and treatments | Partially | Ochiai, CNNFL, FLUCCS (no DNNs) |
| | | | DeepFL, Grace (DNNs) |
| | Response variable, elaboration and metric | Yes | Recall at Top-N, MFR, MAR |
| Docigo | Design type | No | |
| Design | Design type Blocking variables | No Deduced | 2-fold cross validation |
| | Held-constant variables | No | 2 Tota cross validation |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Defects4J (V2.2.0) |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Yes!!!! | Dell, 300 GB RAM, Intel Xeon CPU E5-2680 |
| | | | v4 @ 2.40 GHz, and 8 24GB GPUs of |
| | | | GEForce RTX3090, running Ubuntu 16.04.6 |
| | | | LTS. PyTorch V1.7.1 |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in Table 2 |
| Analysis | Descriptive statistics | Partially | Averages??? (cross validation) |
| | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partially | Missing conclusion |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper A deep learning model for estimating story points

ID AP46 Experiments 7

Comments In RQ3 Deep-SE is not compared. Therefore, I am not counting it

| Aspect | Element | E1 | Comments |
|---------------------|---|--|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 5 | |
| | | #neurons/layer: size of LSTM memory cell= RHWN size of recurrent layer=embedding size | Look-up table for word embeddings. Seen embedding layer is input, but not clear Pre-training is run several times against a validation set and early stopping to choose the best model. Perplexity is used as evaluation metric Issues in training set created before issues validation set, before issues in test set Very well described. Explicitly linked Number of word embeddings dimensions (10, 50, 100, 200) and number of hidden layers in RHN (12 from 2 to 200) |
| | | connections: input, LSTM, average pooling, RHN, feedforward activation functions:linear (feedforward) params. Initialization: Pre-training of embedding and LSTM layers (100 runs and 50 batch size, initial learning rate 0.02, | validation set and early stopping to choose the best model. Perplexity is used as evaluation metric |
| | | adaptation 0.99 and smoothing factor 10 ⁻⁷) | |
| | Model parameters | biases: No weights: No | |
| | DL algorithm | representation: Yes model type: Yes loss function: Difference between predicted and ground truth story points | |
| | Training hyperparameters | regularization: early stopping, dropout (0.5) optimization: RMSprop train-test split: 60-20-20 (creation time) | Issues in training set created before issues in |
| | | learning rate: 0.01 (initial), adaptation 0.9, smoothing 10-6 #iterations: No batch size: 100 #epochs: 1,000 | validation set, before issues in test set |
| | Training data | Yes | Very well described. Explicitly linked |
| Operationalization | Factors and treatments Response variable, elaboration and metric | Yes | (10, 50, 100, 200) and number of hidden layers in RHN (12 from 2 to 200) Mean absolute error, median absolute error |
| Design | Design type | No | |
| Design | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | |
| ocamenation | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Python using Theano |
| | reciniologicariimasuucture | ratuany | MacOS laptop with 2.4GHz Intel Core i5, |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics Inferential statistics | No No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing internal |
| Artifact | Availability | Yes | <u> </u> |
| | Badge | No | |
| | J - | | |

ID AP46 Experiments

In RQ3 Deep-SE is not compared. Ther Comments

| Aspect | Element | E2 |
|---------------------|---|--|
| Experiment type | | Evaluation |
| Hypotheses | Research hypotheses | Yes |
| | Statistical hypotheses | No |
| Variables selection | Model hyperparameters | Same as previous |
| | | |
| | Model parameters | Same as previous |
| | DL algorithm | Same as previous |
| | | |
| | Training hyperparameters | Same as previous |
| Onerationalization | Training data Factors and treatments | Same as previous Factor but not DNN treatment |
| o perationalization | | r dotor but not bring deduncing |
| | Response variable, elaboration and metric | Yes |
| Design | Design type | No |
| DC31B11 | Blocking variables | Project? |
| | Held-constant variables | No |
| | Measured variables (covariates) | No |
| | Randomization | No |
| | Task duration | No |
| | Procedure | No |
| | Number of experimental units | No |
| Instrumentation | Test set | Yes |
| | Measuring instruments | No |
| | Measurement procedure | No |
| | Technological infrastructure | Partially |
| | · · | , |
| Population | Objects (chars. of the experimental datasets) | Yes |
| Analysis | Descriptive statistics | Yes |
| | Inferential statistics | Yes |
| Validity evaluation | Conclusion, internal, construct, external | Partially |
| Artifact | Availability | Yes |
| | Badge | No |
| | | |

ID AP46 Experiments 7

Comments In RQ3 Deep-SE is not compared. Ther

| Aspect | Element | Comments |
|---------------------|------------------------|-------------------------------|
| Experiment type | | |
| Hypotheses | Research hypotheses | RQ1 |
| | Statistical hypotheses | |
| Variables selection | Model hyperparameters | #word embedding dimensions=50 |
| | | #hidden layers=10 |

Model parameters

DL algorithm

Training hyperparameters

| | Training data | |
|---------------------|--|--|
| Operationalization | Factors and treatments | Prediction model (Deep-SE, random guessing, mean effort, median effort) |
| | Response variable, elaboration and metric | Mean absolute error, median absolute error, standardized accuracy, estimated SPs. For Deep-SE Pre-training time, training time, testing time |
| Design | Design type Blocking variables Held-constant variables | |
| | Measured variables (covariates) Randomization | |
| | Task duration Procedure | |
| | Number of experimental units | |
| Instrumentation | Test set | |
| | Measuring instruments | |
| | Measurement procedure | District Theory |
| | Technological infrastructure | Python using Theano MacOS laptop with 2.4GHz Intel Core i5,8GB RAM |
| Population | Objects (chars. of the experimental datasets) | |
| Analysis | Descriptive statistics Inferential statistics | MAE, MeAE and SA are the DS Wilcoxon signed rank test (w Bonferroni correction) and Vargha and Delaney's effect size for estimated SPs |
| Validity evaluation | Conclusion, internal, construct, external | Missing internal |
| Artifact | Availability Badge | |

ID AP46 Experiments 7

Comments In RQ3 Deep-SE is not compared. Ther

| Aspect | Element | E3 | Comments |
|-------------------|--------------------------|------------------|-------------------------------|
| Experiment type | : | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selecti | on Model hyperparameters | Same as previous | #word embedding dimensions=50 |
| | | | #hidden lavers=10 |

Model parameters Same as previous

DL algorithm Same as previous

| | Training data | Same as previous | |
|---------------------|---|------------------------------|--|
| Operationalization | Factors and treatments | Factor but not DNN treatment | Regressor (LSTM+RF, LSTM+SVM, LSTM+ATLM, LSTM+LR, Deep- SE:LSTM+RHN) |
| | Response variable, elaboration and metric | Yes | Mean absolute error, median absolute error, standardized accuracy, estimated SPs |
| Design | Design type | No | |
| · · | Blocking variables | Project? | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Python using Theano MacOS laptop with 2.4GHz Intel Core i5, 8GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Yes | MAE, MeAE and SA are the DS |
| , | Inferential statistics | Yes | Wilcoxon signed rank test (w Bonferroni correction) and Vargha and Delaney's effect size for estimated SPs |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing internal |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP46 Experiments 7

Comments In RQ3 Deep-SE is not compared. Ther

| Aspect | Element | E4 | Comments |
|---------------------|-------------------------|------------------|-------------------------------|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | n Model hyperparameters | Same as previous | #word embedding dimensions=50 |
| | | | #hidden layers=10 |

Model parameters Same as previous

DL algorithm Same as previous

| | Training data | Same as previous | |
|---------------------|---|--------------------------------------|---|
| Operationalization | Factors and treatments | Factor but not DNN treatment | Method (Deep-SE, ABEO) |
| | Response variable, elaboration and metric | Yes | Mean absolute error, estimated SPs |
| Design | Design type | No | |
| Design | Blocking variables | Project, repository (within-between) | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Python using Theano |
| | | | MacOS laptop with 2.4GHz Intel Core i5, |
| | | | 8GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Yes | MAE, MeAE and SA are the DS |
| | Inferential statistics | Yes | Wilcoxon signed rank test (w Bonferroni |
| | | | correction) and Vargha and Delaney's |
| | | | effect size for estimated SPs |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing internal |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP46 Experiments 7

Comments In RQ3 Deep-SE is not compared. Ther

| Aspect | Element | E5 | Comments |
|---------------------|--------------------------|------------------|-------------------------------|
| Experiment type | | Generalization | Adjusted Story points |
| Hypotheses | Research hypotheses | Yes | RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | on Model hyperparameters | Same as previous | #word embedding dimensions=50 |
| | | | #hidden lavers=10 |

Model parameters Same as previous

DL algorithm Same as previous

| | Training data | Same as previous | |
|---------------------|---|------------------------------|---|
| Operationalization | Factors and treatments | Factor but not DNN treatment | Approach (Deep-SE, LSTM+RF, BoW+RF, d2v+RF, LSTM+SVM, LSTM+ATLM, LSTM+LR, mean, median) |
| | Response variable, elaboration and metric | Yes | Mean absolute error, median absolute error, standardized accuracy, estimated adjusted SPs |
| Design | Design type | No | |
| · · | Blocking variables | Project? | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Python using Theano |
| | | | MacOS laptop with 2.4GHz Intel Core i5, |
| | | | 8GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Yes | MAE, MeAE and SA are the DS |
| | Inferential statistics | Yes | Wilcoxon signed rank test (w Bonferroni |
| | | | correction) and Vargha and Delaney's |
| | | | effect size for estimated SPs |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing internal |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP46 Experiments 7

Artifact

Availability

Badge

Comments In RQ3 Deep-SE is not compared. Ther

| Aspect | Element | E6 | Comments |
|---------------------|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ6 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | #word embedding dimensions=50 #hidden layers=10 |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Same as previous Factor but not DNN treatment | Approach (Deep-SE, Porru) |
| | Response variable, elaboration and metric | Yes | Mean absolute error, adjusted SPs |
| Design | Design type | No | |
| - | Blocking variables | Project? | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Python using Theano |
| | . so. morogram minustracture | | MacOS laptop with 2.4GHz Intel Core i5, 8GB RAM |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Yes | MAE, MeAE and SA are the DS |
| | Inferential statistics | Yes | Wilcoxon signed rank test (w Bonferroni |
| | | | correction) and Vargha and Delaney's effect size for estimated SPs |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Missing internal |
| , Craidation | | | |

Yes

No

Paper Semantic learning and emulation based cross-platform binary vulnerability seeker

ID AP50 Experiments 5

Comments Looks like k-fold cross-validation is used for training only (???)

| Aspect | Element | E1 | Comments |
|---------------------------------------|---|--|--|
| Experiment type | | | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: input + 6 hidden + output #neurons/layer: (64? Is embedding size) connections: fully activation functions: Relu, tahn params. Initialization: No | #units input and each hidden layer = #vertices in the original graph Explanations about #units and activation functions are not totally clear The role of n=2 (embedding depth) not clear |
| | | | The fole of h=2 (embedding depth) not clear |
| | Model parameters | biases: No weights:No | |
| | DL algorithm | representation: Yes model type: Yes loss function: Yes regularization: No optimization: No | |
| | Training hyperparameters | train-test split: learning rate: 0.0001 #iterations: No batch size: 10 #epochs: 100 | |
| | Training data | Yes | Datasets from previous studies |
| Operationalization | Factors and treatments | Partially | Tool (BinSeeker, BinSeker-, Genius, Gemini, CA-compare) |
| | Response variable, elaboration and metric | Yes | Accuracy: average ranking where the |
| | response variable, claboration and metre | | vulnerability appears (of 23: optimization |
| Design | Design type | No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, |
| Design | | | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 |
| Design | Design type | No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), |
| Design | Design type Blocking variables | No vulnerability? | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), |
| Design | Design type Blocking variables Held-constant variables | No vulnerability? | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) | No vulnerability? No No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No vulnerability? No No No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units | No vulnerability? No No No No No No No No No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) |
| Design | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No vulnerability? No N | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments | No vulnerability? No N | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set | No vulnerability? No N | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No vulnerability? No N | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for |
| | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure | No vulnerability? No N | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure | No vulnerability? No No No No No No No No Partially | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) | No vulnerability? No No No No No No No No Partially | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA GeForce 1070 GPU, Ubuntu 14.04 LTS |
| Instrumentation | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No vulnerability? No No No No No No No Partially No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA GeForce 1070 GPU, Ubuntu 14.04 LTS |
| Instrumentation Population Analysis | Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics | No vulnerability? No No No No No No No No Partially No No | vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20 MRR 10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2) 10-fold cross-validation?? Different from training set (Dataset II) IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA GeForce 1070 GPU, Ubuntu 14.04 LTS |

ID AP50 Experiments 5

| Aspect | Element | E2 | Comments |
|---------------------|---|------------------|--|
| Experiment type | Describe hunotheses | Vac | PO1 |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | DE digordini | Same as previous | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| On anti-nation | Training data | Same as previous | Tarl (BioCarless Comisi) |
| Operationalization | Factors and treatments | Partially | Tool (BinSeeker- , Gemini) |
| | | | |
| | Response variable, elaboration and metric | Partially | effectiveness (AUC, ROC) |
| | , | | No formula provided |
| | | | |
| Design | Design type | No | |
| | Blocking variables | No | 10-fold cross-validation (as before) |
| | | | |
| | Held-constant variables Measured variables (covariates) | No No | |
| | Randomization | | |
| | Task duration | No No | |
| | Procedure | No | |
| | Number of experimental units | No | 10-fold cross-validation?? |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | Ç |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | IDAPython (create CFG and feature extraction), |
| | | • | LLVM IR plugin. TensorFlow for NN |
| | | | 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA |
| | | | GeForce 1070 GPU, Ubuntu 14.04 LTS |
| | | | |
| Population | Objects (chars. of the experimental datasets) | Same as previous | The DV is accounted to the |
| Analysis | Descriptive statistics | No | The RV is averaged already |
| Validity ovaluation | Inferential statistics Conclusion internal construct external | No Partially | List of throats not grouped in sate series |
| Artifact | Conclusion, internal, construct, external Availability | Partially Yes | List of threats not grouped in categories |
| Ailliact | Badge | No | |
| | 24456 | 110 | |

ID AP50 Experiments 5

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------------------|--|
| Experiment type | | | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | Representation of info (CFG, DFG, PDG,LSFG) |
| | Response variable, elaboration and metric | Partially | effectiveness (AUC, ROC) No formula provided |
| D: | Daving has | NI= | |
| Design | Design type Blocking variables | No No | 10-fold cross-validation (as before) |
| | Held-constant variables Measured variables (covariates) Randomization Task duration Procedure | No No No No | |
| | Number of experimental units | No | 10-fold cross-validation?? |
| Instrumentation | Test set Measuring instruments Measurement procedure Technological infrastructure | Yes No No Partially | IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA GeForce 1070 GPU, Ubuntu 14.04 LTS |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| Analysis | Descriptive statistics | No | The RV is averaged already |
| A 10 10 1 12 12 | Inferential statistics | No | |
| | Conclusion, internal, construct, external | Partially | List of threats not grouped in categories |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP50 Experiments 5

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | 7,-,- | | |
| | | | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Partially | Tool (BinSeeker, BinSeker-, Genius, |
| | | | Gemini, CA-compare) |
| | Response variable, elaboration and metric | Yes | Search time, training time (in seconds) |
| | | | |
| Design | Design type | No | |
| | Blocking variables | No | |
| | | | |
| | Held-constant variables | Deduced | X86-GCC-O0 version used only |
| | Measured variables (covariates) | No | · |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | 10-fold cross-validation?? |
| Instrumentation | Test set | Yes | Dataset I and Dataset II |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | IDAPython (create CFG and feature extraction), LLVM IR plugin. TensorFlow for NN |
| | | | 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA GeForce 1070 GPU, Ubuntu 14.04 LTS |
| Population | Objects (chars. of the experimental datasets) | Same as previous | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | List of threats not grouped in categories |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP50 Experiments 5

| Aspect | Element | E5 | Comments |
|---------------------|---|------------------------------|---|
| Experiment type | | Hyper-parameters fine-tuning | |
| Hypotheses | Research hypotheses | Yes | No |
| • • | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| | Training data | Same as previous | |
| Operationalization | Factors and treatments | Yes | E5: #training epochs (1100), approach (BinSeeker, Gemini) E6: embedding size p (16,64,128,256,Gemini) E7: embedding depth n (15, Gemini) E8: iterations T (1,2,4,6,8, Gemini) |
| | Response variable, elaboration and metric | Yes | E5: loss, AUC E6-E8: ROC, AUC |
| Design | Design trung | Ne | |
| Design | Design type Blocking variables | No No | |
| | Held-constant variables Measured variables (covariates) Randomization | No No No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | 10-fold cross-validation?? |
| Instrumentation | Test set | Yes | Dataset I |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | IDAPython (create CFG and feature extraction), |
| | | | LLVM IR plugin. TensorFlow for NN |
| | | | 8-core 3.60GHz Intel i7, 8GB rAM, NVIDIA |
| | | | GeForce 1070 GPU, Ubuntu 14.04 LTS |
| Population | Objects (chars. of the experimental datasets) | Same as previous | The DV is accounted by |
| Analysis | Descriptive statistics | No | The RV is averaged already |
| Validity evaluation | Inferential statistics Conclusion, internal, construct, external | No Partially | List of threats not grouped in categories |
| Artifact | Availability | Yes | Est of directo not grouped in categories |
| | Badge | No | |
| | - | | |

Paper Easy-to-deploy API extraction by multi-level feature embedding and transfer learning

ID AP51 Experiments 3

Element

Badge

Aspect

Comments E1 fine-tuning. There are 2 more experiments, not related to the DNN, but to transfer learning

E1

Comments

| Aspect | Element | E1 | Comments |
|---------------------|---|--------------|---|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | No | |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as next | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as next | |
| | DL algorithm | Same as next | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as next | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Explained how it was obtained and dropbox |
| | | | link |
| Operationalization | Factors and treatments | Yes | CNN number of filters (20, 40, 50, 80, 100) |
| | | | Dimensions of word embeddings (50, 100, |
| | | | 200, 400) |
| | Response variable, elaboration and metric | No | Not mentioned |
| Design | Design type | No | Not mentioned |
| Design | Blocking variables | Deduced | Library (matalotlib numby pandas openal |
| | blocking variables | Deduced | Library (matplotlib, numpy, pandas, opengl, JDBC, react) |
| | Hald as not out on taking | Ne | JDBC, react) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Same as training. Explained how it was |
| | | | obtained and dropbox link |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |

No

Paper Easy-to-deploy API extraction by multi

ID AP51 Experiments 3

Comments E1 fine-tuning. There are 2 more expe

| Aspect | Element | E2 | Comments |
|---------------------|---|--|--|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| variables selection | Model hyperparameters | #layers: C1: 3, C3: 1, C4:1 #neurons/layer: C1: input (92), filter window size= 3, number of filters=40, word embeddings=200, hidden LSTM units=50, output vector=100 connections: C1: input, convolutional, max-pooling; C3: bidirectional LSTM; C4: softmax activation functions: C1: relu | Has 4 components: C1) char-level features, C2) word embedding (GloVe), C3) sentence-context features, C4) softmax. It mentions the whole DNN is trained end-to-end, but it seems GloVe is trained separately |
| | Model parameters | biases: No weights: No | |
| | DL algorithm | representation: Yes model type: Yes loss function: No regularization: Dropout=0.5 (output of BLSTM) | Mentions loss of word embeddings only |
| | Training hyperparameters | optimization: Adam train-test split: 60-20-20 learning rate: No #iterations: No batch size: No #epochs: 40 | It seems there is one different train per library |
| | Training data | Yes | Explained how it was obtained and dropbox link |
| Operationalization | Factors and treatments | Partially | Model (approach, basic CRF, full CRF). CRF is machine learning (not DNN) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type Blocking variables | No Deduced | Library (matplotlib, numpy, pandas, opengl, JDBC, react) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training. Explained how it was obtained and dropbox link |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Easy-to-deploy API extraction by multi

ID AP51 Experiments 3

Aspect

Comments E1 fine-tuning. There are 2 more expe

Element

Badge

| Aspect | Element | E3 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Optimization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | Woder parameters | Same as previous | |
| | DI algorithm | Como os provious | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | Explained how it was obtained and |
| | Training data | 163 | • |
| 0 | Fortered | De attell | dropbox link |
| Operationalization | Factors and treatments | Partially | Model (complete, without CNN, without |
| | | | word embeddings, witout Bi-LSTM) |
| | | | |
| | Response variable, elaboration and metric | Yes | Precision, recall, F1 |
| Design | Design type | No | |
| | Blocking variables | Deduced | Library (matplotlib, numpy, pandas, |
| | | | opengl, JDBC, react) |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training. Explained how it was |
| | | | obtained and dropbox link |
| | Measuring instruments | No | p |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | No | |
| and ly 313 | Inferential statistics | No | |
| /alidity evaluation | | Partially | Flat list |
| | Conclusion, internal, construct, external | • | riatiist |
| Artifact | Availability | Yes | |
| | Banue | IMO | |

No

E3

ID AP52 Experiments 6

Comments Might be hidden ones. RQ3 corresponds to DNN with inputs only numbers (classified as ML not DL)

| Aspect | Element | E1 | Comments |
|---------------------|---|-------------------------------------|--|
| Experiment type | | Comparison+optimization | Envy detection |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ2, RQ7, RQ8 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 4x2+3. | |
| | | #neurons/layer: Input1: 200x5x3; | |
| | | CNN1: 128 filters, kernel size=1; | |
| | | input2:2(?); CNN2=CNN1; Flatten1:?; | |
| | | Flatten2:?; Merge:?; Dense:128; | |
| | | Output:2 | |
| | | connections: Yes | |
| | | activation functions: CNNs: tanh, | |
| | | rest:? | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: binary cross entropy | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Training is done in 9 out of 10 applications, |
| | | learning rate: No | testing is done in the remaining application |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Linked |
| Operationalization | Factors and treatments | Partially | Model type (approach with/without |
| | | | bootstrap aggregating, JDeodorant) |
| | Response variable, elaboration and metric | Partially | Precision, recall, F1, MCC, AUC, accuracy, |
| | | | time (for approach). Precision, recall, F1 are |
| | | | not defined |
| Design | Design type | No | |
| | Blocking variables | Deduced | Application |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure Number of experimental units | No No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. The remaining application |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| | Technological infrastructure | Partially | Hardware but not software |
| Population | Objects (chars. of the experimental datasets) | Yes | naraware but not software |
| Analysis | Descriptive statistics | Partially | Average only |
| , 5.5 | Inferential statistics | No | Comparison made on averages |
| | | | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | | | |

ID AP52 Experiments 6

Comments Might be hidden ones. RQ3 correspon

| Aspect | Element | E2 | Comments |
|---------------------|---|--|---|
| Experiment type | | Comparison+optimization | Large class |
| Hypotheses | Research hypotheses | Yes | RQ4, RQ7, RQ8 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 2x2+3 | |
| | | <pre>#neurons/layer: Input1:200x5x2;</pre> | |
| | | Input2: 12; rest:? | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | | | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Training is done in 9 out of 10 applications, |
| | | learning rate: No | testing is done in the remaining application |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Linked |
| Operationalization | Factors and treatments | Partially | Model type (approach with/without bootstrap |
| | | | aggregating, DECOR) |
| | Response variable, elaboration and metric | Partially | Precision, recall, F1, MCC, AUC, accuracy, time |
| | | | (for approach). Precision, recall, F1 are not |
| | | | defined |
| Design | Design type | No | |
| | Blocking variables | Deduced | Application |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. The remaining application |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| | Technological infrastructure | Partially | Hardware but not software |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average only |
| | Inferential statistics | No | Comparison made on averages |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | | | |

ID AP52 Experiments 6

Comments Might be hidden ones. RQ3 correspon

| Aspect | Element | E3 | Comments |
|------------------------|---|---|--|
| Experiment type | | Comparison+optimization | Misplaced class |
| Hypotheses | Research hypotheses | Yes | RQ5, RQ6, RQ7, RQ8 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 3x2+3 | |
| | | <pre>#neurons/layer: Input1: 200x5x3;</pre> | |
| | | Input2: 8; rest? | |
| | | connections: Yes | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | | | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: No | |
| | | regularization: No | |
| | | optimization: No | |
| | Training hyperparameters | train-test split: Yes | Training is done in 9 out of 10 |
| | | learning rate: No | applications, testing is done in the |
| | | #iterations: No | remaining application |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Linked |
| Operation a lization | Factors and treatments | Partially | Model type (approach with/without |
| | | | bootstrap aggregating, TACO) |
| | Response variable, elaboration and metric | Partially | Precision, recall, F1, MCC, AUC, accuracy, |
| | | | time (for approach). Precision, recall, F1 |
| | | | are not defined |
| Design | Design type | No | |
| | Blocking variables | Deduced | Application |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | C |
| Instrumentation | Number of experimental units | No Vos | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. The remaining application |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| Denulation | Technological infrastructure | Partially | Hardware but not software |
| Population Analysis | Objects (chars. of the experimental datasets) Descriptive statistics | Yes Partially | Average only |
| Alidiysis | • | • | |
| | Inferential statistics | No | Comparison made on averages |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |
| | Badge | No | |

ID AP52 Experiments 6

Comments Might be hidden ones. RQ3 correspon-

| | | E4 | Comments | |
|---------------------|------------------------|----------------|----------------|--|
| Experiment type | | Generalization | Envy detection | |
| Hypotheses | Research hypotheses | Yes | CS-RQ1 | |
| ; | Statistical hypotheses | No | | |
| Variables selection | Model hyperparameters | Same as E1 | | |

Model parameters Same as E1

DL algorithm Same as E1

Training hyperparameters Same as E1 Training is done in all 10 applications used

in E1-E3

| | Training data | Yes | Linked |
|---------------------|---|-----------|--|
| Operationalization | Factors and treatments | Partially | Model type (approach, JDeodorant) |
| | Response variable, elaboration and metric | Partially | #report, #accepted, #accepted targets, precision, accuracy |
| Design | Design type | No | |
| | Blocking variables | Deduced | Application |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. 5 new applications |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| | Technological infrastructure | Partially | Hardware but not software |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average only |
| | Inferential statistics | Yes | 1-way ANOVA |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP52 Experiments 6

Comments Might be hidden ones. RQ3 correspon-

| Aspect | Element | E5 | Comments | |
|---------------------|------------------------|----------------|-------------|--|
| Experiment type | | Generalization | Large class | |
| Hypotheses | Research hypotheses | Yes | CS-RQ1 | |
| | Statistical hypotheses | No | | |
| Variables selection | Model hyperparameters | Same as E2 | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Model parameters Same as E2

DL algorithm Same as E2

Training hyperparameters Same as E2 Training is done in all 10 applications used

in E1-E3

| | Training data | Yes | Linked |
|---------------------|---|-----------|-------------------------------------|
| Operationalization | Factors and treatments | Partially | Model type (approach, DECOR) |
| | Response variable, elaboration and metric | Partially | #report, #accepted, precision |
| Design | Design type | No | |
| J | Blocking variables | Deduced | Application |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. 5 new applications |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| | Technological infrastructure | Partially | Hardware but not software |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average only |
| | Inferential statistics | No | Comparison made on averages. Effect |
| | | | size provided |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID AP52 Experiments 6

Artifact

Availability

Badge

Comments Might be hidden ones. RQ3 correspond

| Aspect | Element | E6 | Comments |
|---------------------|---|------------------|--|
| Experiment type | | Generalization | Misplaced class |
| Hypotheses | Research hypotheses | Yes | CS-RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as E3 | |
| | | | |
| | | | |
| | Model parameters | Same as E3 | |
| | DL algorithm | Same as E3 | |
| | Training hyperparameters | Same as E3 | Training is done in all 10 applications used in E1-E3 |
| Operationalization | Training data Factors and treatments | Yes Partially | Linked Model type (approach, TACO) |
| | Response variable, elaboration and metric | Partially | #report, #accepted, #accepted targets, precision, accuracy |
| Design | Design type | No | |
| Ü | Blocking variables | Deduced | Application |
| | Held-constant variables | No | • • |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | Seems 1 run |
| Instrumentation | Test set | Yes | Linked. 5 new applications |
| | Measuring instruments | Partially | Only for MCC, AUC |
| | Measurement procedure | Partially | Only for MCC, AUC |
| | Technological infrastructure | Partially | Hardware but not software |
| Population | Objects (chars. of the experimental datasets) | Yes | |
| Analysis | Descriptive statistics | Partially | Average only |
| | Inferential statistics | No | Comparison made on averages |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| , | , , | * I | * * * * * |

Yes

No

ID whichvariables

ID AP53 Experiments 4

| Aspect | Element | E1 | Comments |
|---------------------|---|--|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1, RQ5 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 4 (input, GRU, self- attention, output) #neurons/layer: input 100, RNN 128, 256 self-attention connections: Yes activation functions: No | Embeddings made with GloVe (pre-trained wikipedia and Gigaword 5) |
| | | params. Initialization: No | |
| | Model parameters | biases: No weights: No | |
| | DL algorithm | representation: Yes model type: Yes loss function: binary cross entropy regularization: No | |
| | Training hyperparameters | optimization: Adam train-test split: 80-10-10 learning rate: No #iterations: No batch size: 80 | The MAP score of the model is used to select the best one while training |
| | Training data | #epochs: 200 Partially | 9 OS Java projects of Apache Foundations. Prefectly explained, but not explicitly linked |
| Operationalization | Factors and treatments Response variable, elaboration and metric | Model | Random guess, IR-comp, IR-flat, IR-mix (no DNNs) IR-WE, proposal (DNNs) Top-k accuracy, MRR, MAP, time (proposal |
| | | | only) |
| Design | Design type | No | |
| | Blocking variables | Deduced | Projects |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in table |
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | Yes | Effect size is also reported |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

ID whichvariables

Badge

ID AP53 Experiments 4

Comments

| Aspect | Element | E2 | Comments |
|---------------------|---|-------------------------|---------------------------------------|
| Experiment type | | Optimization | Ablation |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | Training hyperparameters | Same as previous | |
| Operationalization | Training data Factors and treatments | Same as previous Model | Ne+RNN+Attn, OE+Attn, OE+RNN, OE+Uni- |
| Operationalization | raciois and treatments | Wodei | RNN+Attn |
| | Response variable, elaboration and metric | Yes | Top-k accuracy, MRR, MAP |
| Design | Design type | No | |
| | Blocking variables | Deduced | Projects |
| | Held-constant variables | No | • |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training |
| | Measuring instruments | No | Ŭ |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in table |
| Analysis | Descriptive statistics | Partially | Average |
| • | Inferential statistics | Yes | Effect size is also reported |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | |
| | • | | |

No

ID whichvariables

ID AP53 Experiments 4

Comments

| Aspect | Element | E3 | Comments | |
|---------------------|------------------------|------------------|----------|--|
| Experiment type | | Generalization | | |
| Hypotheses | Research hypotheses | Yes | RQ3 | |
| | Statistical hypotheses | No | | |
| Variables selection | Model hyperparameters | Same as previous | | |

Model parameters Same as previous

DL algorithm Same as previous

Training hyperparameters Same as previous

Training data Same as previous

| Operationalization | Factors and treatments | Training | Within-cross training |
|---------------------|---|-----------|--------------------------|
| | Response variable, elaboration and metric | Yes | Top-k accuracy, MRR, MAP |
| Design | Design type | No | |
| | Blocking variables | Deduced | Projects |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in table |
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Flat list |
| Artifact | Availability | Yes | _ |
| | Badge | No | |

ID whichvariables

ID AP53 Experiments 4

| Aspect | Element | E4 | Comments |
|---------------------|---|------------------|-------------------------------------|
| Experiment type | | Optimization | Influence of fitness measures while |
| | | | training |
| Hypotheses | Research hypotheses | Yes | RQ4 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Mandal and a service state of | 6 | |
| | Model parameters | Same as previous | |
| | DI algorithm | Cama as provious | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | · | | |
| | | | |
| | | | |
| | | | |
| | Training data | Same as previous | |
| | | | |
| Operationalization | Factors and treatments | Fitness measure | ACC1, ACC2, MRR, MAP |
| | | | |
| | Despense variable elaboration and metric | Yes | Tonk accuracy MPR MAR |
| | Response variable, elaboration and metric | res | Top-k accuracy, MRR, MAP |
| Design | Design type | No | |
| 0 | Blocking variables | Deduced | Projects |
| | Held-constant variables | No | 3 |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Partially | Same as training |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially | Pytorch |
| Population | Objects (chars. of the experimental datasets) | Yes | Described in table |
| Analysis | Descriptive statistics | Partially | Average |
| A . 10 10 | Inferential statistics | No | |
| Validity evaluation | | Partially | Flat list |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Mining fix patterns for findbugs violations

ID AP54 Experiments 2 Comments

| Aspect | Element | E1 | Comments |
|---------------------|--|--------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ5-1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 7 | Connections: Input, subsampling, |
| | | #neurons/layer: 1000 in | convolutional, subsampling, dense, output |
| | | convolutional | |
| | | connections: Yes | Max pool(?) |
| | | activation functions: softmax, | |
| | | leakrelu | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: Mean squared | |
| | | logarithmic error | |
| | | regularization: No | |
| | | optimization: SGD | |
| | Training hyperparameters | train-test split: No | |
| | 5 // 1 | learning rate: 1e-3 | |
| | | #iterations: No | |
| | | batch size: No | |
| | | #epochs: No | |
| | Training data | Yes | Released own dataset |
| Operationalization | Factors and treatments | Model | No comparison assessment |
| , | Response variable, elaboration and metric | Partially | Unfixed violations resolved |
| Design | Design type | No | |
| | Blocking variables | Deduced | Type of violation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| Danielak'an | Technological infrastructure | No | Described |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | No | |
| Validity evaluation | Inferential statistics Conclusion, internal, construct, external | No Partially | Internal and external only |
| Artifact | Availability | Yes | internal and external only |
| Artifact | Badge | No | |
| | Dauge | 110 | |

Paper Mining fix patterns for findbugs violat

ID AP54 Experiments 2

Comments

| Aspect | Element | E2 | Comments |
|---------------------|------------------------|------------------|----------|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ5-2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |

Model parameters Same as previous

DL algorithm Same as previous

| | Training data | Same as previous | |
|---------------------|---|------------------|--------------------------------|
| Operationalization | Factors and treatments | Model | No comparison. Assessment |
| | Response variable, elaboration and metric | Partially | Fixed bugs |
| Design | Design type | No | |
| | Blocking variables | No | |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Defects4J (publicly available) |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | No | |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | No | |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Partially | Internal and external only |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper

Automatic feature learning for predicting vulnerable software components

ID Experiments AP55 3

| Aspect | Element | E1 | Comments |
|---------------------|---|----------------------------------|---|
| Experiment type | | Evaluation | |
| Hypotheses | Research hypotheses | Yes | RQ1 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | #layers: 3(?) | |
| | | #neurons/layer: No | |
| | | connections: LSTM | |
| | | activation functions: No | |
| | | params. Initialization: No | |
| | Model parameters | biases: No | |
| | | weights: No | |
| | DL algorithm | representation: Yes | |
| | | model type: Yes | |
| | | loss function: log-loss (cross | |
| | | entropy) | |
| | | regularization: dropout (0.5) in | |
| | | LSTM layer | |
| | | optimization: SGD, RMSProp | |
| | Training hyperparameters | train-test split: Yes | |
| | | learning rate: 0.02 | |
| | | #iterations: No | |
| | | batch size: 50 | |
| | | #epochs: No | |
| | Training data | Yes | 2 already available datasets |
| Operationalization | Factors and treatments | Partially | Sw metrics, Bag of Words, Deep belief |
| | | | network, proposed approach (3 variants) |
| | Response variable, elaboration and metric | Yes | Precision, recall, F-measure, AUC (from a |
| | | | confusion matrix) |
| Design | Design type | No | |
| | Blocking variables | Deduced | 10 cross-fold validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially (OS and version of | Theano, Keras (Python). Intel® Xeon® CPU |
| | | python) | E5-2670 0 2.6Gh, 2 CPUs (each 8 cores or 16 |
| | | | threads, 128GB RAM) |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | Partially | Average. Boxplot, but not of the RV, but of |
| | | | the difference |
| | Inferential statistics | No | |
| Validity evaluation | Conclusion, internal, construct, external | Yes | They are perfect |
| Artifact | Availability | Yes | |
| | Badge | No | |

Paper Automatic feature learning for predict

ID AP55 Experiments 3

| Aspect | Element | E2 | Comments |
|---------------------|--|----------------------------------|--|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ2 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | Model parameters | Same as previous | |
| | DI algorithm | Samo as provinus | |
| | DL algorithm | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | • | |
| | | | |
| | | | |
| | Training data | Yes | 2 already available datasets |
| Operationalization | Factors and treatments | Approach and Application version | |
| | Response variable, elaboration and metric | Yes | Performance |
| - Davidson | Post of the second seco | N. | |
| Design | Design type | No Dadwood | arana falal validation |
| | Blocking variables | Deduced | cross-fold validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially (OS and version of | Theano, Keras (Python). Intel® Xeon® CPU E5- |
| | | python) | 2670 0 2.6Gh, 2 CPUs (each 8 cores or 16 |
| | | • | threads, 128GB RAM) |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | Partially | Average |
| | Inferential statistics | No | |
| Validity evaluation | | Yes | They are perfect |
| Artifact | Availability | Yes | - A - Co Decision |
| | Badge | No | |

Paper Automatic feature learning for predict

ID AP55 Experiments 3

| _ | |
|--------|-----|
| Commen | ıts |

| Aspect | Element | E3 | Comments |
|---------------------|---|--------------------------------------|--------------------------------------|
| Experiment type | | Generalization | |
| Hypotheses | Research hypotheses | Yes | RQ3 |
| | Statistical hypotheses | No | |
| Variables selection | Model hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Madal navamatara | Como os provious | |
| | Model parameters | Same as previous | |
| | DL algorithm | Same as previous | |
| | DE digonami | Same as previous | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | Training hyperparameters | Same as previous | |
| | | | |
| | | | |
| | | | |
| | Training data | Yes | 2 already available datasets |
| Operationalization | Factors and treatments | Approach and Cross-application | 2 dileddy available datasets |
| Operationalization | ractors and deadnerits | Approach and cross-application | |
| | Response variable, elaboration and metric | Yes | Performance |
| | , | | |
| Design | Design type | No | |
| | Blocking variables | Deduced | cross-fold validation |
| | Held-constant variables | No | |
| | Measured variables (covariates) | No | |
| | Randomization | No | |
| | Task duration | No | |
| | Procedure | No | |
| | Number of experimental units | No | |
| Instrumentation | Test set | Yes | Same as training set |
| | Measuring instruments | No | |
| | Measurement procedure | No | |
| | Technological infrastructure | Partially (OS and version of python) | Theano, Keras (Python). Intel® Xeon® |
| | recimologicariiinastractare | randally (05 and version of python) | CPU E5-2670 0 2.6Gh, 2 CPUs (each 8 |
| | | | cores or 16 threads, 128GB RAM) |
| Population | Objects (chars. of the experimental datasets) | Yes | Described |
| Analysis | Descriptive statistics | Partially | Average |
| Allalysis | 2 cos. pare stationed | . a. aany | |
| | Inferential statistics | No | |
| Validity evaluation | | Yes | They are perfect |
| Artifact | Availability | Yes | |
| | Badge | No | |