Paper	From UI designs image to GUI skeleton: A neural machine translator to bootstrap mobile GUI implementation
ID	AP1
Experiments	4
Comments	The first experiment is model hyperparameters fine-tuning. No info is given about it. Could be embedded in E2 or

Aspect	**Element**	**E1**	**Comments**
Experiment type		Optimization	Model hyperparameters fine-tuning. "Hidden" except beam-width that appears with E2
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 14 (?) connections: (convolutional + pooling)x6 + RNN encoder (LSTM) + RNN decoder #neurons/layer: Not mentioned activation functions: convolutional (ReLU), pooling (?), encoder (?), decoder (?) params. Initialization: No	Hidden states in LSTM=256. Filter size=3, stride=1 and zero pading=2 for conv layers, first conv layer=64 filters (subsequents x2), pooling units=2x2 and stride=2 for pooling layers #CNN layers, #conv layers filers, #RNN hidden states fine-tuned using another randomly selected 3% of dataset
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Partially model type: Yes loss function: Yes regularization: No optimization: No	Unknown size of input
	Training hyperparameters	train-test split:Yes learning rate: No #iterations: No batch size: No #epochs: 10	90%-3%
	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 (1..5)
	Response variable, elaboration and metric	No	Unknown
Design	Design type	No	
	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	
	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 paarallel--each 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
Analysis	Descriptive statistics	No	No results at all are provided from this experiment
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	From UI designs image to GUI skeletc
ID	AP1
Experiments	4
Comments	The first experiment is model hyperpasimply missing. E3 (generalization using other test sets). Just assesses h

Aspect	Element	E2	Comments
Experiment type		Evaluation	Does not really compare against anything
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 14 connections: (convolutional + pooling)x6 + RNN encoder (LSTM) + RNN decoder #neurons/layer: Not mentioned activation functions: convolutional (ReLU), pooling (?), encoder (?), decoder (?) params. Initialization: No	Hidden states in LSTM=256. Filter size=3, stride=1 and zero pading=2 for conv layers, first conv layer=64 filters (subsequents x2), pooling units=2x2 and stride=2 for pooling layers Not sure if these are the definite values obtained fom the previous experiment. Except beam-width that 2 is chosen
	Model 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 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 (3--9), #GUI coponents (5..65 step 5), @containers (3..24 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 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	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	From UI designs image to GUI skeletc
ID	AP1
Experiments	4
Comments	The first experiment is model hyperpaow good it is.

Aspect	Element	E3	Comments
Experiment type		Generalization	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	Same as previous?	Beam-widt=2 is explicitly mentioned
	Model parameters	Same as previous	
	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	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	No	
	Blocking variables	Deduced	Could be app??? 20
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Number of experimental units	No	1??
Instrumentation	Test set	Partially	20 completely unseen apps
	Measuring instruments	Partially	Can be deduced
	Measurement procedure	Partially	Can be deduced
	Pre-processing	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
Analysis	Descriptive statistics	Yes	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	From UI designs image to GUI skeletc
ID	AP1
Experiments	4
Comments	The first experiment is model hyperpa

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 ones)
	Measuring instruments	Partially	Can be deduced
	Measurement procedure	Partially	Can be deduced
	Pre-processing	Yes	
	Dataset construction		
	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
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	Deep code search
ID	AP2
Experiments	1
Comments	

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, MLP L1:100, MLP L3: 400, maxpooling:? connections: Yes activation functions: tahn params. Initialization: No	
	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
	Response variable, elaboration and metric	Yes	FRank, Success-rate@k, Precision@k, MRR. Described in detail
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	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 queries, 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	

Paper	Approach to debug method names based on the analysis of consistency between method names and method code	
ID	AP3	
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) #neurons/layer: 1000 (?). Input layer is nxk k=94 connections: L2, L4 (convolutional), L3, L5 (subsampling, maxpool), L6 (dense) activation functions: softmax (output), ReLU (rest) params. initialization: No	Number of layers taken from figure Table IV. Number of nodes in hidden layers is 1000 but not clear if is total number or per layer(?).
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type (architecture): Yes loss function: mean absolute error regularization: No optimization: SGD	
	Training hyperparameters	train-test split: Different datasets learning rate: 1e-2 #iterations: No batch size: No #epochs: No	Refers to [43] and [60] for parameters used
	Training data	Yes	Described in paper and included in artifact
Operationalization	Factors and treatments	Partially	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	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	Yes	K (size of sets of adjacent vectors). Outside of DNN
	Randomization	No	
	Task duration	No	
	Number of experimental units	No	Looks like 1 run
Instrumentation	Test set	Yes	Built separately from training
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	DL4J library	Implements using paragraph vector, wrod2vec 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	Conclusion, internal, construct, external	Internal, external	
Artifact	Availability	Github	Updated recently!!! (8 months ago)
	Badge	No	

Paper	Approach to debug method names based on th
ID	AP3
Experiments	3
Comments	Preprocessing is done with state-of-the-art NNslo exactly the same task

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	Same as previous	
Operationalization	Factors and treatments	Partially	DNN proposed is assessed, but not compared
	Response variable, elaboration and metric	Inconsistency avoidance (T1), first-token accuracy (T2), full-name accuracy (T3)	Fully defined.
Design	Design type	No	Should be a nested design. 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	
	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	
	Badge	Same as previous	

Paper	Approach to debug method names based on th
ID	AP3
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	Conclusion, internal, construct, external	Same as previous	
Artifact	Availability	Same as previous	
	Badge	Same as previous	

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 units are factors	
	Model parameters	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	

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 evaluated in E1 is finally chosen
		#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	
		learning rate: No	
		#iterations: No	
		batch size: No	
		#epochs: 60k	
	Training data	Yes	Available in artifact
Operationalization	Factors and treatments	Test set	Model is only assessed
	Response variable, elaboration and metric	Yes	Raw count and percentage of successfully predicted code changes
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	

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
	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	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	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	Linked to artifact
	Measuring instruments	No	
	Measurement procedure	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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	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
	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	Linked to artifact
Operationalization	Factors and treatments	DNN Model	DeepTyper (JSNice is not a DNN), NL2Type For DeepTyper, use their publicly available artifact, and do not apply confidence threshold. Same test and training sets are used for both approaches
	Response variable, elaboration and metric	Yes	Precision, recall, F1 on top-1 predicted
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	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	Yes	Linked to artifact
	Measuring instruments	No	
	Measurement procedure	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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	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**	**E3**	**Comments**
Experiment type		Evaluation	Evaluate DNN for another task (inconsistencies detection)
Hypotheses	Research hypotheses	Yes	RQ3
	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	Linked to artifact
Operationalization	Factors and treatments	DNN	1 treatment only
	Response variable, elaboration and metric	Yes	Frequency of potential inconsistency types (inconsistency/non-standard type annotation/misclassification)
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	Partially	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	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	Yes	Linked to artifact
	Measuring instruments	Yes	
	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.	
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	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	E4	Comments
Experiment type		Optimization	It should be the first experiment...
Hypotheses	Research hypotheses	Yes	RQ4
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 4 #neurons/layer: l1 (?), l2 (100?), l3 (256?), l4 (1000?) connections: input, bi-directional LSTM, fully connected, output (softmax) activation functions: No params. Initialization: No	Embedding built upon Word2Vec, apparently being re-trained, 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) #neurons deduced
	Model parameters	biases: No weights: No	But they are in the artefact.
	DL algorithm	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 #epochs: 12	No need of K-cross validation, due to large amount of data
	Training data	Yes	Linked to artifact
Operationalization	Factors and treatments	DNN architecture	output of DNN (5...5000) 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	No	
	Held-constant variables	Partially	Input representation: words in names: 6 words in comment: 12 words in comment: 10 #pars: 10
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	
Instrumentation	Test set	Yes	Linked to artifact
	Measuring instruments	No	
	Measurement procedure	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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

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 vs late fusion. But layers are the same. Late fusion is siamese. Probably #neurons/layer would change.
		#neurons/layer: No	
		connections: No	
		activation functions: No	
		params. Initialization: No	
	Model parameters	biases: No	
		weights: No	
	DL algorithm	representation: Partially	
		model type: Yes	
		loss function: Yes	
		regularization: No	
		optimization: No	
	Training hyperparameters	train-test split: Yes	80%-20%
		learning rate: No	
		#iterations: No	
		batch size: No	
		#epochs: No	
	Training data	Yes	80%
Operationalization	Factors and treatments	Yes	Input change region strategies (change-contrast, action-continuity, both) DNN architecture (early vs late fusion)
	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)	No	
	Randomization	No	
	Task duration	No	
	Procedure	Partially	Each model is trained with same data
	Number of experimental units	No	
Instrumentation	Test set	Yes	20%
	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 per strategy and architecture
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	No	
	Badge	No	

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**
Experiment type		Generalization	Are 3 experiments. Behaviour in other developers/working environmnets/programming languages
Hypotheses	Research hypotheses	Yes	RQ2
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	Best from factors in previous experiment	both and early-fusion
	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	
Instrumentation	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	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	No	
	Badge	No	

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
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	No	late-fusion. Does not mention anything else
	Model parameters	No	
	DL algorithm	No	
	Training hyperparameters	No	
	Training data	No	
Operationalization	Factors and treatments	Yes	1 treatment DNN
	Response variable, elaboration and metric	Partially	Time
Design	Design type	Partially	Deduced
	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	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	
	Badge	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 learning rate: No #iterations: No batch size: No #epochs: No	Different sets are used
	Training data	Same as next	
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)
	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	Conclusion, internal, construct, external	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	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 Randoop
	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	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	
Artifact	Availability	Yes	
	Badge	No	

Paper	A novel neural source code representation based on abstract syntax tree
ID	AP8
Experiments	4
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	Source code classification
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 6? #neurons/layer: GRU (100) connections: encoder, recurrent, pooling, output activation functions: Some mentioned: identity (encoder) params. Initialization: No	For encodder mentions Word2Vec. Its role not explained. Could be the pre-trained encoder (mesning weights are initialized with these values)
	Model parameters	biases: Yes weights: Yes	Explicitly says that "trained models are stored"
	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
Operationalization	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
Design	Design type	No	Seems 1 factor-6 treatment (TextCNN, LSTM, TBCNN,LSCNN,PGD+GGNN)
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure		
	Number of experimental units	No	Assume 1
Instrumentation	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
Population	Objects (chars. of the experimental dataset)	Partially	Mention the datasets and references (OJ)
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

Paper	A novel neural source code represent
ID	AP8
Experiments	4
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
	Training hyperparameters	Same as previous	#epochs: max. 5 Threshold: 0.5
	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	
Population	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
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 novel neural source code represent
ID	AP8
Experiments	4
Comments	

Aspect	Element	E3	Comments
Experiment type		Optimization	Several architectural choices
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	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	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified
Artifact	Availability	Yes	
	Badge	Yes	Available badge

Paper	A novel neural source code represent
ID	AP8
Experiments	4
Comments	

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
ID	AP9
Experiments	2
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	
Hypotheses	Research hypotheses	Yes	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: Yes #neurons/layer: Partially connections: Yes activation functions: No params. Initialization: No	5 layers: input (x3), Embedding (x3), GRU(x3), Attention (x2), Dense (1) Input size: 100, 123, 100 Embedding size: ? GRU sizes: 256 each All this is deduced from text. Other numbers are given, but it is not straightforward to see what they are
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: No regularization: No optimization: No	The missing info can be found in the code. This is a good example of code and paper being inconsistent!!!!
	Training hyperparameters	train-test split: Yes learning rate: #iterations: batch size: #epochs: 10	90%-5%-5% All DNNs 10 epochs
	Training data	Yes	95% dataset.
Operationalization	Factors and treatments	Partially	DNN: attendgru, SBT, codenn, ast-attendgru 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	BLEU1..4, 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?
Instrumentation	Test set	Yes	5% dataset
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	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.
Population	Objects (chars. of the experimental datasets)	Partially	Java methods from the Sourcerer repository provided by Lopes et al.
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Present a list of 2 threats. No division
Artifact	Availability	Yes	
	Badge	No	

Paper	A neural model for generating natural
ID	AP9
Experiments	2
Comments	

Aspect	Element	E2	Comments
Experiment type		Optimization	Assuming only AST available (no internal documentation)
Hypotheses	Research hypotheses	Yes	
	Statistical hypotheses	No	
Variables 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
	Response variable, elaboration and metric	Partially	BLEU1..4, 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?
Instrumentation	Test set	Same as previous?	Code removed
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	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.
Population	Objects (chars. of the experimental datasets)	Partially	Java methods from the Sourcerer repository provided by Lopes et al.
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Present a list of 2 threats. No division
Artifact	Availability	Yes	
	Badge	No	

Paper	DeepPerf: performance prediction for configurable software with deep sparse neural network
ID	AP10
Experiments	4
Comments	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: l1 (), l2..n+1(128?), ln+2(1) connections: activation functions: params. Initialization: Xavier (weights)	
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes 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: learning rate: initial between 0.0001-0.1, dropped by 0.001 #iterations: batch size: Size of training data #epochs: 2000?	
	Training data	Yes	Input and output are normalized (0-1 and 0-100). 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	No	
	Procedure	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	
	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 publications where they are fully explained
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	E1 are a series of experiments. Difficu

Aspect	Element	E2	Comments
Experiment type		Evaluation	
Hypotheses	Research hypotheses	Yes	RQ1, 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	
	Training data	Yes	Input and output are normalized (0-1 and 0-100)
Operationalization	Factors and treatments	Model Type, Subject system (?)	DECART, DeepPerf (DECART is classification trees) Not sure if subject system (apache, x264,BDB-J, LLVM, BDB-C, SQLite). Could be blocking variable
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time
Design	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	
	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	
	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 publications where they are fully explained
Analysis	Descriptive statistics	Yes	Mean and 95% CI
	Inferential statistics	Yes	t-test
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	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	
	Training data	Yes	Input and output are normalized (0-1 and 0-100)
Operationalization	Factors and treatments	DNN architecture	SPLConqueror (no DNN), DeepPerf (DECART is classification trees)
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time
Design	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	
	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	
	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 publications where they are fully explained
Analysis	Descriptive statistics	Yes	Mean and 95% CI
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	E1 are a series of experiments. Difficu

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	
	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	
	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	
	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 publications where they are fully explained
Analysis	Descriptive statistics	Yes	Mean and 95%CI
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	Unblind your apps: Predicting natural-language labels for mobile gui components by deep learning
ID	AP11
Experiments	2
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: Difficult #neurons/layer: No connections: Difficult activation functions: Some params. Initialization: No	Resnet-101 Architecture pre-trained on MS COCO dataset as CNN module (removing last pooling layer)
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: Relative entropy regularization: No optimization: Adam for encoder/decoder	beta1, beta2 and epsilon given for Adam Loss function not clear if whole net or encoder/decoder only
	Training hyperparameters	train-test split: Yes learning rate: Variable #iterations:No batch size:No #epochs:No	80% training, 10% test, 10% validation (for each app. category) Formula for learning rate given
	Training data	Yes	Linked to github
Operationalization	Factors and treatments	Neural models	CNN+LSTM, CNN+CNN, LabelDroid The other two approaches do not seem to be from SE proposals, but ideas from authors. Nothing is said about other approaches
	Response variable, elaboration and metric	Partial	Missing description of metric
Design	Design type	No	
	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	
Artifact	Availability	Yes	https://github.com/chenjshnn/LabelDroid
	Badge	No	

Paper	Unblind your apps: Predicting natural-
ID	AP11
Experiments	2
Comments	

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	
	Response variable, elaboration and metric	Yes	
Design	Design type	No	
	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	Yes	5-point Likert scale
	Measurement procedure	Yes	Expert assessment
	Technological infrastructure	Missing OS	Same as before
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	Yes	Mean and std (per app type and total)
	Inferential statistics	Yes	Seems incorrect Wilcoxon (should be Friedman) Not explained why non-parametrics Not clear which are the datapoints
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	Same as before
	Badge	No	

Paper	CC2Vec: Distributed representations of code changes
ID:	AP12
Experiments	4
Comments	

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
		model type: Yes	in regularization and optimization
		loss function: Yes	described?
		regularization: Yes	
		optimization: Adam	
	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
Instrumentation	Test set	Partially	Mentions whole dataset
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	No	
Population	Objects (chars. of the experimental datasets)	No	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity 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	
Validity evaluation	Conclusion, internal, construct, external	Internal, external	
Artifact	Availability	Yes	
	Badge	No	

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.
	Training data	Same as previous	Mentions whole dataset
Operationalization	Factors and treatments	Neural model	CC2Vec used to improve SOTA DeepJIT
	Response variable, elaboration and metric	Yes	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	
Validity evaluation	Conclusion, internal, construct, external	Internal, external	
Artifact	Availability	Yes	
	Badge	No	

Paper	CC2Vec: Distributed representations o
ID:	AP12
Experiments	4
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	
	Training data	Parially	The 3 datasets of previous experiments
Operationalization	Factors and treatments	Comparison function (7 levels)	Different comparison functions in the comparison layers
	Response variable, elaboration and metric	Yes	BLEU-4, F1, AUC
Design	Design type	No	
	Blocking variables	Yes	Dataset?
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Number of experimental units	No	Nothing mentioned now about 5-fold
Instrumentation	Test set	Partially	Same as previous
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	No	
Population	Objects (chars. of the experimental datasets)	Partially	Same as previous
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Internal, external	
Artifact	Availability	Yes	
	Badge	No	

Paper	Software visualization and deep transfer learning for effective software defect prediction	
ID	AP13	
Experiments	5	Although the first one could be 3
Comments		

Aspect	**Element**	**E1**	**Comments**
Experiment type		Optimization	Different hyperparameters
Hypotheses	Research hypotheses	No	No RQ
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 10 #neurons/layer: No connections: Partially activation functions: Partially params. Initialization: Partially	Uses as base AlexNet network structure Coneections described for non-AlexNet layers Params initialization is given only for AlexNet (as it is pre-trained) Some activation functions only (perhaps not all layers have activation function)
	Model parameters	biases: No weights: No	Not mentioned in the paper
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: Yes optimization: No	Data augmentation is used for regularization
	Training hyperparameters	train-test split:Yes learning rate: No #iterations: No batch size: No #epochs: 500	AlexNet is pre-trained with ImageNet 2012
	Training data	yes	Explicitly linked
Operationalization	Factors and treatments	hyperparms (levels given)	Not sure if could be 3 experiments (with 2 factors) 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	
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**	**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, DTL-DP	The other models are never explained. Mentions 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
Comments	

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?	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 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	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 GPUs
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
Artifact	Availability	Yes	
	Badge	No	

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)
	Training data	Yes	Explicitly linked
Operationalization	Factors and treatments	DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP)	
	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	Factors and treatments	DTL-DP architecture (Base, +TL, +Atten, +Aug, DTL-DP)	
	Response variable, elaboration and metric	Yes	Fmeasure, time
Design	Design type	No	
	Blocking variables	No	
	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	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 model type: Yes loss function: No regularization: No optimization: No	
	Training hyperparameters	train-test split: Yes learning rate: Yes #iterations: No batch size: No #epochs: Yes	Mentions "epoch size" I understand this is #epochs (not batch size)
	Training data	Yes	Data linked to artifact
Operationalization	Factors and treatments	word2vec vector length (100, 50, 120), learning rate (0.001, 0.005, 0.01), epoch size (100, 200, 300)	Beam search used
	Response variable, elaboration and metric	Yes	Top K (k=1,5,10)
Design	Design type	No	
	Blocking variables	Yes	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 (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
	Training data	Yes	Data linked to artifact
Operationalization	Factors and treatments	Neural model (Ratchet, Tufano(18), CODIT, Tufano(19))	Do not mention anything about them. Jus that 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:	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. 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	

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 (non-siamse version)	
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 2+4(x2)+1+1 #neurons/layer: l1, l2 (512) connections: l1, l2 (feedforward), l3 (input), l4 (convolutional), l5 (BiLSTM), l6 (combination), l7 (similarity) activation functions: l1, l2 softsign params. Initialization: l1, l2 (trained)	2 first layers are SOTA disentanglement (already trained). Not clear if this is approach or not. I would say not, as it is not trained. All descriptions are partial. Grid search used for: POS tag embedding (50), kernel sizes (2,3,4,5) feature 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 weights: No	Although l1 and l2 corresponds to a SOTA disentanglement NN, and they are available
	DL algorithm	representation: Partially model type: Partially loss function: Cross-entropy regularization: Dropout (0.1) and early stopping (after 10 epochs) optimization: No	It is not clear if their approach includes disentanglement or not. They mention dropout and early stopping as optimization, but it seems to me they are regularization
	Training hyperparameters	train-test split: Yes learning rate: No #iterations: No batch size: No #epochs: No	3-fold intra-project-cross-validation from 3 projects
	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	
	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	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 said
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 f
ID	AP15
Experiments	3
Comments	Not sure the extent of the proposed s(

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?	
	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	
	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
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	
Population	Objects (chars. of the experimental datasets)	Same as previous	
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 f
ID	AP15
Experiments	3
Comments	Not sure the extent of the proposed s(

Aspect	**Element**	**E3**	**Comments**
Experiment type		Comparison with SOTA and optimization (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 #epochs: No	3-fold cross-project-cross-validation from 3 projects
	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	
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	RetrievaL-based neural source cocde summarization	
ID	AP16	
Experiments	5	E5 is difficult to assess with this template. E4 encompasses 2 RQs
Comments	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	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: Yes #neurons/layer: No connections: Yes activation functions: No params. Initialization: No	hidden states in LSTM=several
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: No optimization: No	No menciona regularization. Lambda = several beam size = several
	Training hyperparameters	train-test split: Yes learning rate: No #iterations: No batch size: No #epochs: No	
	Training data	Yes	
Operationalization	Factors and treatments	Factors: embedding size (256), hidden states=512, batch size=32, maximum iterations=100k, Adam optimizer, learning rate=0.001, beam size=5, lambda=3	Factors can be deduced, treatments are not explained "selecting the best one among some alternatives"
	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
	Technological infrastructure	Yes	Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory
Population	Objects (chars. of the experimental datasets)	Yes	
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	RetrievaL-based neural source cocde s
ID	AP16
Experiments	5
Comments	Some "hidden" experiments. Experim

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 #neurons/layer: Yes connections: Yes activation functions: No params. Initialization: No	Embedding size=256 hidden states in LSTM=512
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: No optimization: Adam	No menciona regularization. Lambda = 3 beam size = 5
	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-Codesum, Hybrid-DRL, GRNMET, Rencos)	Does not mention anything about the others. Proposal is partially explained
	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
	Technological infrastructure	Yes	Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory
Population	Objects (chars. of the experimental datasets)	Yes	
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	RetrievaL-based neural source cocde s
ID	AP16
Experiments	5
Comments	Some "hidden" experiments. Experim

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	
	DL algorithm	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	
	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	Partially	For number of low-frequency words correctly generated it is partially explained
	Technological infrastructure	Yes	Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory
Population	Objects (chars. of the experimental datasets)	Yes	
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	RetrievaL-based neural source cocde s
ID	AP16
Experiments	5
Comments	Some "hidden" experiments. Experim

Aspect	**Element**	**E4**	**Comments**
Experiment type		Optimization	Components not part of DNN
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	
Operationalization	Factors and treatments	Yes k (1..10), T (0, 0.2, 0.5, 0.8)	
	Response variable, elaboration and metric	Yes	AVERAGED BLEU, 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	
	Measurement procedure	No	
	Technological infrastructure	Yes	Implemented using OpenNMT. Ubuntu 16.04 server with 16 cores of 2.4GHz CPU, 128Gb RAM, Titan Xp GPU with 12GB memory
Population	Objects (chars. of the experimental datasets)	Yes	
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	RetrievaL-based neural source cocde s
ID	AP16
Experiments	5
Comments	Some "hidden" experiments. Experim

Aspect	**Element**	**E5**	**Comments**
Experiment type		Evaluation	Humans assess similarity between proposed and real comment
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	Neural model (Hybrid-DRL, Rencos)	
	Response variable, elaboration and metric	Yes	Similarity as Likert scale (1-5)
Design	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
Instrumentation	Test set	Partially	Do not mention which samples chosen. People from Amazon MT
	Measuring instruments	Yes	
	Measurement procedure	Yes	
	Technological infrastructure	Yes	
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	Yes	
	Inferential statistics	Partially	Not sure if correct. Wilcoxon
Validity evaluation	Conclusion, internal, construct, external	Partially	Present a list of threats. No division
Artifact	Availability	Yes	
	Badge	No	

Paper	On learning meaningul assert statements for unit test cases
ID	AP17
Experiments	3
Comments	E1 is hidden

Aspect	**Element**	**E1**	**Comments**
Experiment type		Optimization	
Hypotheses	Research hypotheses	Yes	RQ1
	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	Raw source and abstracted code
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	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	
	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**	**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 type: Yes loss function: negative log likelihood regularization: encoder & decoder dropout (0.2) optimization: Adam	Model is encoder/attention/decoder
	Training hyperparameters	train-test split: 80-10-10 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, complementariness raw/abstract models, # and % asserts resolved with copy mechanism, time to generated assert statements
Design	Design type	No	
	Blocking variables	Yes	Beam size (1..50 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	
	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	
Design	Design type	No	
	Blocking variables	Yes	Beam size (1, 5, 10)
	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	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 #neurons/layer: 3 convolution 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	GloVe for word embeddings and implement TextCNN for sentence encoding
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes 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 learning rate: No #iterations: No batch size: No #epochs: No	Grid search for hyper-parameter tuning
	Training data	Yes	Linked
Operationalization	Factors and treatments	Partially	Model type. The others are not SOTA, but created by the authors
	Response variable, elaboration and metric	Yes	With formulas
Design	Design type	No	
	Blocking variables	Deduced	Programming community
	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
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Partially	Keras, 2.5GHz Intel Core i5, 8GB DDR3 RAM
Population	Objects (chars. of the experimental datasets)	Yes	Explained
Analysis	Descriptive statistics	Partially	Average
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, construct, external. Missing conclusion
Artifact	Availability	Yes	
	Badge	No	

Paper	Automatically matching bug reports with related app reviews
ID	AP19
Experiments	1
Comments	

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 #neurons/layer: No connections: No activation functions: No params. Initialization: No	DISTILBERT creates 768-dimensional embeddings. Nothing is mention on whether defaults are used for its implementation (layers, units, etc.) or something is changed
	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	**Element**	**E1**	**Comments**
Experiment type		Optimization	
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 #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	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	
Validity evaluation	Conclusion, internal, construct, external	Partially	Named limitations. Flat list
Artifact	Availability	Yes	Github
	Badge	No	

Paper	CURE: Code-aware neural machine tra
ID	AP20
Experiments	3
Comments	E1 is hidden. They acknowledge to haain 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 weights: No	
	DL algorithm	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 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	
	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	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 #neurons/layer: Some connections: Yes activation functions: Some params. Initialization: No	Generic description of the structure of the layers, but no exact number of encoders/decoders ReLU for output layer of last encoder
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Transformer loss function: No regularization: No optimization: Adam	BPE (byte pari enconding. Do not mention if a DNN is used for this)
	Training hyperparameters	train-test split: 80-10-10 learning rate: 0.001 #iterations: No batch size: 256 #epochs: 147	Beam search=10 Length normalization parameter alpha=0.6 Grid search for hp tuning
	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	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	Fault localization with code coverage representation learning
ID	AP22
Experiments	4
Comments	R3 and RQ4 are outside DNN

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	Conclusion, internal, construct, external	Partially	Flat list of 2 threats
Artifact	Availability	Yes	
	Badge	No	

Paper	Fault localization with code coverage
ID	AP22
Experiments	4
Comments	R3 and RQ4 are outside DNN

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	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	Conclusion, internal, construct, external	Partially	Flat list of 2 threats
Artifact	Availability	Yes	
	Badge	No	

Paper	Fault localization with code coverage
ID	AP22
Experiments	4
Comments	R3 and RQ4 are outside DNN

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	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	Conclusion, internal, construct, external	Partially	Flat list of 2 threats
Artifact	Availability	Yes	
	Badge	No	

Paper	Fault localization with code coverage
ID	AP22
Experiments	4
Comments	R3 and RQ4 are outside DNN

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	
Operationalization	Factors and treatments	Partially	Does not compare against SOTA. Merely assesses
	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)
	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	

Paper	Fault localization with code coverage
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	
	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	

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 #neurons/layer: No connections: Yes activation functions: No params. Initialization: Random	Pytorch implementation of GPT-2. 6 transformer blocks, six heads per block, n_ctx =1000, embedding_dim = 300
	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 learning rate: 1e-3 #iterations: No batch size: No #epochs: Yes (for all models)	"Other hyperparameters" borrowed from [20]
	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, model size, MRR
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	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

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	
Operationalization	Factors and treatments	Deep3, Code2Seq, PathTrans, 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)	No	
	Randomization	No	
	Task duration	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 e

Aspect	**Element**	**E3**	**Comments**
Experiment type		Optimization	
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	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)	No	
	Randomization	No	
	Task duration	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 e

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	
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	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	A context-based automated approach for method name consistency checking and suggestion
ID	AP24
Experiments	6
Comments	

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	
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**	**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
Comments	

Aspect	**Element**	**E3, E4**	**Comments**
Experiment type		Optimization	Two tasks
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	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
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	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	No	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
	Response variable, elaboration and metric	Partially	Testing errors and F-score
Design	Design type	No	
	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	No	
	Number of experimental units	Yes	10 (10-fold cross-validation)
Instrumentation	Test set	No	
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Yes	
Population	Objects (chars. of the experimental datasets)	No	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	DeepSim: Deep learning code functior
ID	AP25
Experiments	3
Comments	E1 is hidden

Aspect	**Element**	**E2**	**Comments**
Experiment type		Evaluation	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: Not clear #neurons/layer: 88-6, (128x6-256-64)-128-32 activation functions: ELU params. Initialization: No	For all models: layers size
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: cross-entropy regularization: L2(0.00003) optimization: dropout (0.75)	For all models, L2 value & dropout
	Training hyperparameters	train-test split: 10-fold-cross.val learning rate: 0.001 (initial) #iterations: No batch size: No #epochs: 4	For all models, learning rate and epochs
	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 functior
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 #neurons/layer: 88-6, (128x6-256-64)-128-32 activation functions: ELU params. Initialization: No	For all models: layers size
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: cross-entropy regularization: L2(0.00003) optimization: dropout (0.75)	For all models, L2 value & dropout
	Training hyperparameters	train-test split: 10-fold-cross.val learning rate: 0.001 (initial) #iterations: No batch size: No #epochs: 4	For all models, learning rate and epochs
	Training data	Yes	Perfect!!! BigCloneBench. Two trainings, full set and set with functionality id=4 only
Operationalization	Factors and treatments	Partially	Model: DECKARD, RtvNN, CDLH (for the 3, values reported in paper), DeepSim
	Response variable, elaboration and metric	Partially	Recall, precision, F1 and per clone type
Design	Design type	No	
	Blocking variables	Partially	For DeepSim. For the paper models, no n-fold cross validation
	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. Reported result is averaged. For RtVNN same full dataset
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)
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	Deep learning type inference
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	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 7 (?) #neurons/layer:n 300 (embedding), 650 (both bi-GRU), ? (rest) connections: ? (Input), embedding, bi-directional GRU, concatenation, bi-directional GRU, ? (projection), softmax, (?) output activation functions: No params. Initialization: No	Not clear n. Layers (missing from Figure 2 input and softmax) Not clear if 2 bi-GRU layers, or they are the same
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: cross-entropy regularization: dropout prob. of 50% (second hidden layer), layer normalization optimization: Adam, momentum	Model is selected when error is stabilized Momentum =1/e after the first 1,000 minibatches
	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 #epochs: 10	
	Training data	Partially???	1,000 top starred open-source projects on Github Feb 28,2018. Available. Not sure if reproducible. "Predominantly consisted of TypeScript code" What is predominantly?
Operationalization	Factors and treatments	No	Unknown which hyper-parameters were fine-tuned and how
	Response variable, elaboration and metric	Partially	Validation 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	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	

Paper	Deep learning type inference
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	

Paper	Deep learning type inference
ID	AP26
Experiments	4
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-DNN), Hybrid
	Response variable, elaboration and metric	Yes	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
Instrumentation	Test set	No	Same as previous?
	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	Deep learning type inference
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 instantiation)+output #neurons/layer: Varies acording to instantiation connections: varies according to instantiation activation functions: relu params. Initialization: No	- CNN: kernel sizes (1,3,5,7), kernel counts (8,16,32,64), max pooling - RNN(LSTM): hidden size(128) - 1-layer perceptron (64) - Multilayer perceptron (16,1)
	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 learning rate: No #iterations: No batch size: No #epochs: No	Training set: 105,108 links, testing set: 43,680 (may) links
	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
Instrumentation	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 GPU (trained & tested on GPU)
Population	Objects (chars. of the experimental datasets)	No	PRIMO
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 2 threats. No link to types
Artifact	Availability	No	
	Badge	No	

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 model type: No loss function: No regularization: No optimization: No	
	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 [1..5]
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	
Validity evaluation	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
Operationalization	Factors and treatments	Partial	FRAFT, Word2Vec, D2Vec
	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
	Technological infrastructure	No	
Population	Objects (chars. of the experimental datasets)	Partial	Some statistics are given (table 6)
Analysis	Descriptive statistics	Partial	Only mean
	Inferential statistics	No	
Validity evaluation	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 f

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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partial	Plain list and very weak
Artifact	Availability	No	
	Badge	No	

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 model type: Yes loss function: sigmoidal cross-entropy regularization: No optimization: dropout (en LSTM), Adam	Input word embedding vectors trained using GloVe
	Training hyperparameters	train-test split: Yes learning rate: 0.001 #iterations: No batch size: 32 #epochs: 100	10-fold cross-validation using 10% of dataset as test set
	Training data	Yes (paper mentions code and data of the study are shared	CADO. Resampling is made to improve it
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 Test set	Glove is trained on 4 corpora for RNN
	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	
Validity evaluation	Conclusion, internal, construct, external	Partially	Not linked in al cases to classification
Artifact	Availability	Yes	
	Badge	Yes	

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 embedding layer and LSTM hidden layer	Grid search is used
	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 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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	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	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	Refers to architecture	multi-layer spec and neural architecture (n-gram, seq2seq, hybrid)
	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	No	
	Procedure	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
Population	Objects (chars. of the experimental datasets)	No	Reference to paper
Analysis	Descriptive statistics	Yes	Mean and std. Dev for ranking
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	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
	Statistical hypotheses	No	
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 weights:No	
	DL algorithm	Same as previous	
	Training hyperparameters	Same as previous	
	Training data	Missing	Same as previous???
Operationalization	Factors and treatments	Refers to architecture	multi-layer spec and neural architecture (n-gram, seq2seq, hybrid)
	Response variable, elaboration and metric	Yes	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
Experiments	2
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 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 paper for HDFS. No info about MS due to confidentiality	Test sets synthetic HDFS dataset (NewTesting1, NewTesting2), Microsoft 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 datsets
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		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	
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 #neurons/layer: No connections: Yes activation functions: No params. Initialization: No	skip-gram for word embeddings Replace 64-dimension softmax layer of DeepMoji with an n-dimension softmax layer
	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	DNN is pre-trained in a different task, later fine-tuned. They build SEntiMoji upon DeepMoji
	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	
	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	

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	
	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	

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	
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	Linked to repo
Operationalization	Factors and treatments	Partially	Train test size
	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	

Paper	TypeWriter: natural type prediction with search -based validation
ID	AP33
Experiments	3
Comments	One more experiment about issues outside DNN

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	SOTA
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: #neurons/layer: 200LSTM connections: 3x1bi-directional RNN based on LSTM+1 hidden layer activation functions: ouput layer (softmax) params. Initialization: No	Word2vec for embeddings. Trained 2 times: code+identifiers and comments
	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: 80-20 learning rate: No #iterations: No batch size 0.005 #epochs: 10	The entire neural model is learned jointly (word2vec?) 2 separate trainings fr argument types and function types Split is by file
	Training data	Partially	One of the 2 datasets is private (Facebook). Other available for download
Operationalization	Factors and treatments	Partially	Model: naïve, DeepTyper (re-implementation), NL2Type (re-implementation)
	Response variable, elaboration and metric	Yes	Precision, recall, F1 (weighted), top-k scores (1,3,5)
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	Test set	Partially	One of the 2 datasets is private (Facebook). Other available for download
	Preprocessing	Yes	Word2vec using gensim, LibCST
	Measuring instruments	Deduced	
	Measurement procedure	Deduced	
	Technological infrastructure	Partially	Python
Population	Objects (chars. of the experimental datasets)	Partially	For public dataset
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Partially	Independent re-implementation. La original es de Facebook
	Badge	No	

Paper	TypeWriter: natural type prediction wi
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	
	Training data	Same as previous	
Operationalization	Factors and treatments	Partially	Model: TypeWriter variants-removing parts (full, 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	Test set	Partially	One of the 2 datasets is private (Facebook). Other available for download
	Preprocessing	Yes	Word2vec using gensim, LibCST
	Measuring instruments	Deduced	
	Measurement procedure	Deduced	
	Technological infrastructure	Partially	Python
Population	Objects (chars. of the experimental datasets)	Partially	For public dataset
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Partially	Independent re-implementation. La original es de Facebook
	Badge	No	

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
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Partially	Independent re-implementation
	Badge	No	

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-time pooling layer+fully connected	Convolutional layer uses 3 sets of convolution kernels with 3 different widhts (3,4,5), with 100 kernels each
		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 set: incidents 01/01/17 to 08/31/2018. Testing set: incidents 09/01/2018 to 10/31/2018. Validation test is 5% training set
		learning rate: No	
		#iterations: No	
		batch size: No	
		#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	
	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	Conclusion, internal, construct, external	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	
	Training data	Same as previous	Privacy issues (Microsoft)
Operationalization	Factors and treatments	Partially	Methods: DWEN (DNN), LIDAR (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	
	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	Conclusion, internal, construct, external	No	
Artifact	Availability	No	
	Badge	No	

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 model type: bidirectional GRU loss function: No regularization: ReLU on fully connected layer optimization: No	Early stop
	Training hyperparameters	train-test split: 80-10-10 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 model type: bidirectional GRU loss function: No regularization: ReLU on fully connected layer optimization: No	Early stop
	Training hyperparameters	train-test split: 80-10-10 learning rate: No #iterations: No batch size: 32 #epochs: No	
	Training data	Yes	Aunque no sé si repetible
Operationalization	Factors and treatments	DNN model	Los treatments no parecen SOTA. Rule-based baseline, language model baseline (RNNLM, RNNLM+PA(1),RNNLM+PA(1-5)), sequence-to-sequence (S2S, S2S+PA(1), S2S+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	
	Number of experimental units	Yes	3 (random seed, different training/validation/test split)
Instrumentation	Test set	Yes	
	Measuring instruments	Yes	
	Measurement procedure	Deduced	
	Technological infrastructure	Partially	Pytorch, OpenNMT
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	Partially	Solo mean
	Inferential statistics	Yes	
Validity evaluation	Conclusion, internal, construct, external	Partially	
Artifact	Availability	Yes	
	Badge	No	

Paper	MTFuzz: Fuzzing with a multi-task neural network
ID	AP36
Experiments	4
Comments	

Aspect	Element	E1	Comments
Experiment type		Evaluation	
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 7 + 3(with 3 paralells) #neurons/layer: L1(?), L2(2048), L3(1024), L$(512) connections: 3 encoder, 3 (x3) decoder activation functions: ReLu for hidden, sigmoid output params. Initialization:	Each task has the same weight
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: multi-task regularization: No optimization: Adam	Loss function: MSE for edge coverage, adaptive loss for edge and context-sensitive edge
	Training hyperparameters	train-test split: No learning rate: 0.001 #iterations: No batch size: No #epochs: 100	750 input samples for re-training
	Training data	Yes	
Operationalization	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	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
Instrumentation	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
Population	Objects (chars. of the experimental datasets)	Partially	Nothing for synthetic bugs
Analysis	Descriptive statistics	Partially	For edge coverage mean and std. Dev.
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	3 threats not classified
Artifact	Availability	Yes	
	Badge	Yes	

Paper	MTFuzz: Fuzzing with a multi-task neu
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	No	
	Measured variables (covariates)	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	Conclusion, internal, construct, external	Partially	3 threats not classified
Artifact	Availability	Yes	
	Badge	Yes	

Paper	MTFuzz: Fuzzing with a multi-task neı
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	
	Blocking variables	Program	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	Not specified this time
	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	Conclusion, internal, construct, external	Partially	3 threats not classified
Artifact	Availability	Yes	
	Badge	Yes	

Paper	MTFuzz: Fuzzing with a multi-task neı
ID	AP36
Experiments	4
Comments	

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 previous	
	Training data	Same as previous	
Operationalization	Factors and treatments	Yes	Program type: ELF files, XML files, Fuzzer (Neuzz, MTFuzz, AFL, MTFuzz inputs+embeddings)
	Response variable, elaboration and metric	Yes	Edge coverage
Design	Design type	No	
	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 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.
	Training data	Same as previous	
Operationalization	Factors and treatments	Model	ACETON w/wo attention
	Response variable, elaboration and metric	Partially	Precision and recall per category. Only names Performance: Time (training and prediction) and F1 per 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	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	
	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	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	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
	Training data	Same as previous	
Operationalization	Factors and treatments	Mobile device (how+sw)	Nexus 5X (Android 7.0) vs Nexus 6P (Android 6.01.1)
	Response variable, elaboration and metric	Partially	Precision, recall
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	
	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	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	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 learning rate: No #iterations: 160 batch size: 8 #epochs: No	5-fold cross-validation
	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	
	Procedure	No	
	Number of experimental units	No	
Instrumentation	Test set	Partially	Train 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	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	Object detection for graphical user int
ID	AP38
Experiments	7
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	
	DL algorithm	Same as previous	
	Training hyperparameters	Same as previous	
	Training data	Same as previous	
Operationalization	Factors and treatments	Approaches. Factor but not treatments	REMAUI, Xianyu, Faster RCNN, YOLOv3, CenterNet. None of them specified
	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	

Paper	Object detection for graphical user int
ID	AP38
Experiments	7
Comments	

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	
	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	
	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
Comments	

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	
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	

Paper	Object detection for graphical user int
ID	AP38
Experiments	7
Comments	

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, YOLOv3,CenterNet, Our method), Element (nontext-only, mix nontext, text both)
	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	

Paper	Object detection for graphical user int
ID	AP38
Experiments	7
Comments	

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	
	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
Comments	

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, Centernet, our method)
	Response variable, elaboration and metric	yes	#bbbox, accuracy, 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	No	
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	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	No	

Paper	STATEFORMER: Fine-grained type recovery from binaries using generative state modeling
ID	AP39
Experiments	8
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	Compares against nothing
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: Some #neurons/layer: No connections: Yes activation functions: No params. Initialization: No	Architecture is described at a high level. For details points to supplementary material
	Model parameters	biases: No weights: No	Does not explicitly mention parameters
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: No optimization: No	Loss: MSE + BCE. Points to supplementary material
	Training hyperparameters	train-test split: 80-10-10 learning rate: No #iterations: No batch size: No #epochs: 10, 50	Pretrain+train Points to supplementary material
	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 (Fairseq toolkit) 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	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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	
	Training data	Same as previous???	Not clear
Operationalization	Factors and treatments	Model type (STATEFORMER)	For EKLAVIA, numbers reported in paper are used.
	Response variable, elaboration and metric	Yes	Accuracy
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	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
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	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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	No	
	Blocking variables	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	Test set	Partially	Mention evaluated on "their" projects
	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	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

Aspect	**Element**	**E5**	**Comments**
Experiment type		Evaluation	Against SOTA
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???	Not clear
Operationalization	Factors and treatments	Model type (STATEFORMER, Debin, Ghidra)	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	
	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	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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	
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	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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	
	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	
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	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	8
Comments	

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	
	Training data	Same as previous???	Not clear
Operationalization	Factors and treatments	Pre-training (no, STATEFORMER, TREX)	TREX is DNN, but they do not mention where they take it from
	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	
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
Experiments	5
Comments	

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 model type: Yes loss function: Yes regularization: Dropout 0.1 optimization: Adam	Loss function: maximize negative log-likelihood of the oracle edit sequence
	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 treatments	jGenProg, HDRepair, Nopol, CapGen, 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 prograı
ID	AP40
Experiments	5
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	
	DL algorithm	same as previous	
	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 prograı
ID	AP40
Experiments	5
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	
	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 prograı
ID	AP40
Experiments	5
Comments	

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	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	
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 progra
ID	AP40
Experiments	5
Comments	

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	
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal
Artifact	Availability	Yes	
	Badge	Available	

Paper	Lightweight global and local contexts guided method name recommendation with prior knowledge
ID	AP41
Experiments	7
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Optimization	Hidden
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	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	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	Lightweight global and local contexts guided me
ID	AP41
Experiments	7
Comments	

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 model type: Yes loss function: Yes regularization: No optimization: No	"Due to page limit, we only briefly introduce this model in the paper, and more details could be referred to the existing work [57]" Loss function:negative log likelihood of the oracle word for that step
	Training hyperparameters	train-test split: Yes learning rate: No #iterations: No batch size: No #epochs: No	
	Training data	Yes	References known datasets (Java-small, Java-med, Java-large, Mnire) publicly available
Operationalization	Factors and treatments	Model type (10 approaches vs Cognac)	For Mnire, they use results reported in paper. Do 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	
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	Lightweight global and local contexts guided me
ID	AP41
Experiments	7
Comments	

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	
	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

Paper	Lightweight global and local contexts guided me
ID	AP41
Experiments	7
Comments	

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

Paper	Lightweight global and local contexts guided me
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	
	Badge	Yes	Available

Paper	Lightweight global and local contexts guided me
ID	AP41
Experiments	7
Comments	

Aspect	**Element**	**E6**	**Comments**
Experiment type		Optimization	use of caller/calle info.
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 dataset (Mnire), publicly available
Operationalization	Factors and treatments	Model type (seq2seq model vs. Cognac	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
Artifact	Availability	Yes	
	Badge	Yes	Available

Paper	Lightweight global and local contexts guided me
ID	AP41
Experiments	7
Comments	

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 #neurons/layer: embeddings: 768x3, rest ??? connections: Yes activation functions: ReLu params. Initialization:	Plus 3 encoders (TODO comment, code change, commit message) that generate embeddings using pre-trained BERT. The 3 encoders are jointly trained with the DNN.
	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 learning rate: 0.001 #iterations: 1,000 batch size: 32 #epochs: No	
	Training data	Yes	Included in RP
Operationalization	Factors and treatments	Partially	Compares against SOTA (TCO, TMO, TCMO, IRSC, 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	
	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	
	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	
	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**	**E3**	**Comments**
Experiment type		Optimization	Replaces BERT with Word2Vec
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	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	
	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	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 i

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 learning rate: No #iterations: No batch size: No #epochs: No	Divide 80-10-10 vulnerable method. For training, the same number of non-vulnerable methods are added. For validation/testing, real ratio between vulnerable/not vulnerable used
	Training data	Partially	Referenced (datasets used by others) but not explicitly linked to artifact (see data/results of experiments). Fan et al, Reveal, FFMPeg+Qemu
Operationalization	Factors and treatments	Partially	Model type (VulDeePecker, Devign, SyseVR, Russel, Reveal, IVDetect). All DNNs
	Response variable, elaboration and metric	Yes	Mean average precision, normalized DCG, first ranking, accuracy under curve, precision, recall, F-score, training and prediction time for IVDetect
Design	Design type	No	
	Blocking variables	Deduced	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	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-grair
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 #iterations: No batch size: No #epochs: No	Training: Reveal, FFMPeg+Qemu, 20% Fan, testing: 80% fan
	Training data	Same as previous	
Operationalization	Factors and treatments	Partially	training type (within/cross) on IVDetect
	Response variable, elaboration and metric	Yes	Mean average precision, normalized DCG
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	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	
Artifact	Availability	Yes	
	Badge	No	

Paper	Vulnerability detection with fine-grair
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	
	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	
	Badge	No	

Paper	Vulnerability detection with fine-grair
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	
	Model parameters	Same as previous	
	DL algorithm	Same as previous	
	Training hyperparameters	Same as E1	
	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	
	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	
	Badge	No	

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	Partially	COALA, LabelDroid. In theory, COALA should be (almost) fully defined in RP, but LabelDroid is not
	Response variable, elaboration and metric	Partially	BLEU, METEOR, ROUGH, CIDEr, exact match, time (for COALA only). Formulas missing
Design	Design type	No	
	Blocking variables	Deduced	Random split (5 times)
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	Yes	NLGE Python library
	Number of experimental units	No	Seems 1 run (per random split)
Instrumentation	Test set	Yes	Same as training
	Measuring instruments	Yes	NLGE Python library
	Measurement procedure	Deduced	
	Technological infrastructure	Partially	PyTorch. Ubuntu wiht NVIDIA GP102 GPU and 128GB memory
Population	Objects (chars. of the experimental datasets)	Yes	Details shown
Analysis	Descriptive statistics	Partially	Value used is mean(?) due to 5 times
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list
Artifact	Availability	Yes	
	Badge	No	

Paper	Boosting coverage-based fault localization via graph-based representation learning
ID	AP45
Experiments	4
Comments	

Aspect	**Element**	**E1**	**Comments**
Experiment type		Evaluation	
Hypotheses	Research hypotheses	Yes	RQ1
	Statistical hypotheses	No	
Variables selection	Model hyperparameters	#layers: 8(?) Embedding, GGNN (5 layers?), fully (?), softmax	Mention layer normalizaion and residual connection
		#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	
	Blocking variables	Deduced	leave-one-out cross validation
	Held-constant variables	No	Fixed random seeds
	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	
	Badge	No	

Paper	Boosting coverage-based fault localiz
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	Partially	Averages??? (cross validation)
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Missing conclusion
Artifact	Availability	Yes	
	Badge	No	

Paper	Boosting coverage-based fault localiza
ID	AP45
Experiments	4
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	
	Training data	Yes	Defects4J (V1.2.0). Publicly available
Operationalization	Factors and treatments	Partially	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	
	Badge	No	

Paper	Boosting coverage-based fault localiza
ID	AP45
Experiments	4
Comments	

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 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
Design	Design type	No	
	Blocking variables	Deduced	2-fold 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	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	
Validity evaluation	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 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, adaptation 0.99 and smoothing factor 10^{-7})	Look-up table for word embeddings. Seems 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
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: Difference between predicted and ground truth story points regularization: early stopping, dropout (0.5) optimization: RMSprop	
	Training hyperparameters	train-test split: 60-20-20 (creation time) learning rate: 0.01 (initial), adaptation 0.9, smoothing 10-6 #iterations: No batch size: 100 #epochs: 1,000	Issues in training set created before issues in validation set, before issues in test set
	Training data	Yes	Very well described. Explicitly linked
Operationalization	Factors and treatments	Yes	Number of word embeddings dimensions (10, 50, 100, 200) and number of hidden layers in RHN (12 from 2 to 200)
	Response variable, elaboration and metric	Yes	Mean absolute error, median absolute error and standardized accuracy
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	
	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	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	Missing internal
Artifact	Availability	Yes	
	Badge	No	

Paper	A deep learning model for estimating
ID	AP46
Experiments	7
Comments	In RQ3 Deep-SE is not compared. Ther

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
	Training data	Same as previous
Operationalization	Factors and treatments	Factor but not DNN treatment
	Response variable, elaboration and metric	Yes
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
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

Paper	A deep learning model for estimating
ID	AP46
Experiments	7
Comments	In RQ3 Deep-SE is not compared. Thei

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	
	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	MAE, MeAE and SA are the DS
	Inferential statistics	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	

Paper	A deep learning model for estimating
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 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	
	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	

Paper	A deep learning model for estimating
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	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	
	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	
	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	

Paper	A deep learning model for estimating
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	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	
	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	

Paper	A deep learning model for estimating
ID	AP46
Experiments	7
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	
	Training data	Same as previous	
Operationalization	Factors and treatments	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 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	

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
	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 vulnerability appears (of 23: optimization level, architecture, compiler), % top-1,3,5,20, MRR
Design	Design type	No	
	Blocking variables	vulnerability?	10-fold cross-validation. But it is weird (not sure they are really using it) Optimization level (x3), architecture (x3), compiler (x2)
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	10-fold cross-validation??
Instrumentation	Test set	Yes	Different from training set (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)	Partially	
Analysis	Descriptive statistics	No	The RV is averaged already
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	List of threats not grouped in categories
Artifact	Availability	Yes	
	Badge	No	

Paper	Semantic learning and emulation bas(
ID	AP50
Experiments	5
Comments	Looks like k-fold cross-validation is us(

Aspect	**Element**	**E2**	**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	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	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	10-fold cross-validation??
Instrumentation	Test set	Yes	Same as training set
	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	The RV is averaged already
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	List of threats not grouped in categories
Artifact	Availability	Yes	
	Badge	No	

Paper	Semantic learning and emulation base
ID	AP50
Experiments	5
Comments	Looks like k-fold cross-validation is use

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
Design	Design type	No	
	Blocking variables	No	10-fold cross-validation (as before)
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	10-fold cross-validation??
Instrumentation	Test set	Yes	Same as training set
	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	The RV is averaged already
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	List of threats not grouped in categories
Artifact	Availability	Yes	
	Badge	No	

Paper	Semantic learning and emulation base
ID	AP50
Experiments	5
Comments	Looks like k-fold cross-validation is us

Aspect	**Element**	**E4**	**Comments**
Experiment type			
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	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	

Paper	Semantic learning and emulation bas
ID	AP50
Experiments	5
Comments	Looks like k-fold cross-validation is us

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 (1…100), approach (BinSeeker, Gemini) E6: embedding size p (16,64,128,256,Gemini) E7: embedding depth n (1…5, 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 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	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	
Analysis	Descriptive statistics	No	The RV is averaged already
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Partially	List of threats not grouped in categories
Artifact	Availability	Yes	
	Badge	No	

Paper	Easy-to-deploy API extraction by multi-level feature embedding and transfer learning
ID	AP51
Experiments	3
Comments	E1 fine-tuning. There are 2 more experiments, not related to the DNN, but to transfer learning

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	
	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	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	
	Badge	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) optimization: Adam	Mentions loss of word embeddings only
	Training hyperparameters	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	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	
	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
Comments	E1 fine-tuning. There are 2 more expe

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	Yes	Explained how it was obtained and 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	
	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	Deep learning based code smell detection
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 learning rate: No #iterations: No batch size: No #epochs: No	Training is done in 9 out of 10 applications, testing is done in the remaining application
	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	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	

Paper	Deep learning based code smell detec
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 #neurons/layer: Input1:200x5x2; 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 learning rate: No #iterations: No batch size: No #epochs: No	Training is done in 9 out of 10 applications, testing is done in the remaining application
	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	

Paper	Deep learning based code smell detec
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 #neurons/layer: Input1: 200x5x3; 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 learning rate: No #iterations: No batch size: No #epochs: No	Training is done in 9 out of 10 applications, testing is done in the remaining application
	Training data	Yes	Linked
Operationalization	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	
	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	

Paper	Deep learning based code smell detec
ID	AP52
Experiments	6
Comments	Might be hidden ones. RQ3 correspon

Aspect	Element	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	

Paper	Deep learning based code smell detec
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	
	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	

Paper	Deep learning based code smell detec
ID	AP52
Experiments	6
Comments	Might be hidden ones. RQ3 correspon

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
	Training data	Yes	Linked
Operationalization	Factors and treatments	Partially	Model type (approach, TACO)
	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	No	Comparison made on averages
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list
Artifact	Availability	Yes	
	Badge	No	

Paper	Which variables should I log?
ID	whichvariables
ID	AP53
Experiments	4
Comments	

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 params. Initialization: No	Embeddings made with GloVe (pre-trained wikipedia and Gigaword 5)
	Model parameters	biases: No weights: No	
	DL algorithm	representation: Yes model type: Yes loss function: binary cross entropy regularization: No optimization: Adam	
	Training hyperparameters	train-test split: 80-10-10 learning rate: No #iterations: No batch size: 80 #epochs: 200	The MAP score of the model is used to select the best one while training
	Training data	Partially	9 OS Java projects of Apache Foundations. Prefectly explained, but not explicitly linked
Operationalization	Factors and treatments	Model	Random guess, IR-comp, IR-flat, IR-mix (no DNNs) IR-WE, proposal (DNNs)
	Response variable, elaboration and metric	Yes	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	

Paper	Which variables should I log?
ID	whichvariables
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	
	Training data	Same as previous	
Operationalization	Factors and treatments	Model	Ne+RNN+Attn, OE+Attn, OE+RNN, OE+Uni-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	
	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	

Paper	Which variables should I log?
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	

Paper	Which variables should I log?
ID	whichvariables
ID	AP53
Experiments	4
Comments	

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	
	Model parameters	Same as previous	
	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
	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	

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 #neurons/layer: 1000 in convolutional connections: Yes activation functions: softmax, leakrelu params. Initialization: No	Connections: Input, subsampling, convolutional, subsampling, dense, output Max pool(?)
	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 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	
	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	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 hyperparameters	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	AP55
Experiments	3
Comments	

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 python)	Theano, Keras (Python). Intel® Xeon® 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. 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
Comments	

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	
	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
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® 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
	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
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	Yes	2 already available datasets
Operationalization	Factors and treatments	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® 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
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	Yes	They are perfect
Artifact	Availability	Yes	
	Badge	No	