NL2Type: Inferring JavaScript function types from natural language information AP5 $\,\,$ 4 $\,\,$ 2 non-comparative experiments (RQ1 and RQ5) and a qualitative (RQ3) Paper ID

Experiments

Aspect Experiment type	Element	E1 Optimization+Evaluation	Comments Not sure what it is	Code
Experiment type		Optimization+Evaluation	Not sure what it is	
Hypotheses	Research hypotheses Statistical hypotheses	Yes No	RQ1, RQ5	
Variables selection	Model hyperparameters	#layers: 3 #neurons/layer: I1 (100), I2 (256), I3 (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)	File "model.h5" provided. When loaded in python: #llayers: 3 connections: input, bidirectional LSTM, dense #neurons/layer: 100_512.978 Activation functions and params. Initialization unknown
	Model parameters	biases: No	#neurons deduced But they are in the artefact.	<u>Yes</u>
	DL algorithm	weights: No representation: Yes model type: Yes loss function: categorical cross entropy regularization: dropout (20%) optimization: Adam		<u>Code not provided</u>
	Training hyperparameters	(defaults?) 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	<u>Code not provided</u>
	Training data	Yes	Linked to artifact	Provided (dropbox link)
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 Blocking variables Held-constant variables	No No No		
	Measured variables (covariates) Randomization Task duration Procedure	No No No No		
	Number of experimental units Test set	No Yes	Seems 1 run	Test est appointed of the sale on the C
Instrumentation	Measuring instruments Measurement procedure	Yes No No	Linked to artifact	Test set provided (dropbox link)
	Technological infrastructure	Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory.		
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries	
Analysis	Descriptive statistics Inferential statistics	No No		
Validity evaluation		No		

NL2Type: Inferring JavaScript function APS 4 2 non-comparative experiments (RQ1

Experiments

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

Paper ID NL2Type: Inferring JavaScript function AP5 4

Experiments

2 non-comparative experiments (RQ1

Aspect	Element	E3	Comments	Code
operiment type		Evaluation	Evaluate DNN for another task (inconsistencies	
			detection)	
potheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No 12		
anables selection	Model hyperparameters	Same as next?		Same as previous
	Model parameters DL algorithm	Same as next? Same as next?		Same as previous Same as previous
	Training hyperparameters	Same as next?		Same as previous
	Training data	Yes	Linked to artifact	Same as previous
perationalization	Factors and treatments	DNN	1 treatment only	
esign	Response variable, elaboration and metric Design type Blocking variables Held-constant variables	Yes No No Partially	Frequency of potential inconsistency types (inconsistency/non-standard type annotation/misclassification)	
	Measured variables (covariates) Randomization Task duration Procedure	No No No Partially	Multiple runs to check the predictions. Some neurons are purposefully deactivated during prediction.	
	Randomization Task duration Procedure	No No Partially	are purposefully deactivated during prediction.	
strumentation	Randomization Task duration	No No		Same as previous
strumentation	Randomization Task duration Procedure Number of experimental units	No No Partially	are purposefully deactivated during prediction. Seems more than 1	Same as previous
strumentation	Randomization Task duration Procedure Number of experimental units Test set Measuring instruments	No No Partially No Yes Yes	are purposefully deactivated during prediction. Seems more than 1 Linked to artifact NL2Type is used in a different way. The return value is used to check if the predicted type matches the real	Same as previous
strumentation	Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No Partially No Yes Yes Yes Yes Yes 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.	are purposefully deactivated during prediction. Seems more than 1 Linked to artifact NL2Type is used in a different way. The return value is used to check if the predicted type matches the real one.	Same as previous
opulation	Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets)	No No Partially No Yes Yes Yes Yes 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.	are purposefully deactivated during prediction. Seems more than 1 Linked to artifact NL2Type is used in a different way. The return value is used to check if the predicted type matches the real	Same as previous
	Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No Partially No Yes Yes Yes Yes Yes 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.	are purposefully deactivated during prediction. Seems more than 1 Linked to artifact NL2Type is used in a different way. The return value is used to check if the predicted type matches the real one.	Same as previous

NL2Type: Inferring JavaScript function AP5 4 2 non-comparative experiments (RQ1 Paper ID

Experiments

Aspect	Element	E4	Comments	Code
Experiment type		Optimization	It should be the first experiment	
Hypotheses	Research hypotheses Statistical hypotheses	Yes No	RQ4	
Variables selection	Model hyperparameters	Hlayers: 4 #neurons/layer: I1 (?), I2 (100?), I3 (256?), I4 (1000?) connections: input, bi-directional LSTM, fully connected, output (softmax) activation functions: No params. Initialization: No	Embedding built upon Word2Vec, apparently being 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	Same as previous
	Model parameters	biases: No weights: No	But they are in the artefact.	Same as previous
	DL algorithm	representation: Yes model type: Yes loss function: categorical cross entropy regularization: dropout (20%) optimization: Adam (defaults?)		Same as previous
	Training hyperparameters	train-test split: 80-20 learning rate: No #ifterations: No batch size: 256 #epochs: 12	No need of K-cross validation, due to large amount of data	
Operationalizati	Training data Factors and treatments	Yes DNN architecture	Linked to artifact	Same as previous
Operationalization	ractors and treatments	DNN architecture	output of DNN (55000) Paper mentions they have run experiments to choose hyperparameters, but they are not described	
	Response variable, elaboration and metric	Yes	Precision, recall, F1 on top-1 predicted	
Design	Design type Blocking variables Held-constant variables	No No Partially	Input representation:	
	neto-constant variables	ratually	words in names: 6 words in comment: 12 words in comment: 10 thore: 10	
	Measured variables (covariates) Randomization Task duration Procedure	No No No No	words in comment: 12	
	Measured variables (covariates) Randomization Task duration Procedure	No No No No	words in comment: 12 words in comment: 10	
Instrumentation	Measured variables (covariates) Randomization Task duration	No No No	words in comment: 12 words in comment: 10	Same as previous
Instrumentation	Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments	No No No No Ves No	words in comment: 12 words in comment: 10 #pars: 10	Same as previous
	Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No No No No 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.	words in comment: 12 words in comment: 10 #pars: 10 Linked to artifact	Same as previous
Instrumentation Population Analysis	Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure	No No No No No 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.	words in comment: 12 words in comment: 10 #pars: 10	Same as previous

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

Experiments

Aspect	Element	E1	Comments	Code
Experiment type		Evaluation	Source code classification	
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 6?	For encodder mentions Word2Vec. Its role not explained.	#neurons/layer: GRU (100)
		#neurons/layer: GRU (100)	Could be the pre-trained encoder (mesning weights are	
		connections: encoder, recurrent,	initialized with these values)	Encode dimension: 128
		pooling, output		
		activation functions: Some		Activation function: relu
		mentioned: identity (encoder)		
		params. Initialization: No		
	Model parameters	biases: Yes	Explicitly says that "trained models are stored"	Cannot see them
		weights: Yes		
	DL algorithm	representation: Yes		loss function: cross-entropy but also
		model type: Yes		BCELoss
		loss function: cross-entropy		regularization: dropout (0.2)
		regularization: No		optimization: AdaMax
		optimization: AdaMax		
	Training hyperparameters	train-test split: 60-20-20		train-test split: not clear. Not
		learning rate: 0.002		understandable formula
		#iterations: No		learning rate: Not found
		batch size: 64		#iterations: No
		#epochs: max. 15		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,	
	m. 1:		TBCNN,LSCNN,PGD+GGNN)	
	Blocking variables	No No		
	Held-constant variables 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		
	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified	
Artifact	Availability	Yes	A the late	
	Badge	Yes	Available	

Paper A novel neural source code represent ID AP8

ID AP8 Experiments 4

Training data

Operationalization Factors and treatments

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

Model parameters

Same as previous?

DL algorithm

Same as previous? Loss function is binary cross-entropy

Same as previous

Training hyperparameters

Same as previous? #epochs: max. 5
Threshold: 0.5

OJ, BCB, referenced ASTNN, RAE+, CDLH

Yes

For other approahces: RAE+: Configuration as in paper CDLH: Not public, results from paper Response variable, elaboration and metric Design type Precision, recall, F1
Seems 1 factor-4 treatment (RAE+,CDLH,PGD+GGNN, ASTNN) Yes Design Blocking variables Held-constant variables Measured variables (covariates) Randomization No No No No Task duration Procedure No No Number of experimental units
Test set
Measuring instruments No Yes Yes Assume 1 OJ, BCB, referenced Instrumentation Measurement procedure 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
Validity evaluation Conclusion, internal, construct, external Partially Only 3 threats are listed, not classified

Artifact Availability
Badge Yes Available badge

Paper ID A novel neural source code represent

AP8 4 Experiments

Aspect	Element	E3	Comments	Code
Experiment type		Optimization	Several architectural choices	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		Same as previous

Model parameters Same as previous? Same as previous DL algorithm Same as previous? Same as previous

Training hyperparameters Same as previous? 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 Used in process and state No Only 3 threats are listed, not classified Yes Available badge

A novel neural source code represent AP8 4 Paper ID

Experiments

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

Model parameters Same as previous? Same as previous DL algorithm Same as previous? Same as previous

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	
	Disables contables	No	
	Blocking variables Held-constant variables	No 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	_	
	Badge	Yes	Available badge

DeepPerf: performance prediction for configurable software with deep sparse neural network AP10 4 E1 are a series of experiments. Difficult to assess how many, as they are described at a very high level

Paper ID Experiments Comments

Aspect	Element	E1	Comments	Code
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		Matches and completes
		#neurons/layer: l1 (), l2n+1(128?),		
		In+2(1)		
		connections:		
		activation functions:		
		params. Initialization: Xavier (weights)		
	Model parameters	biases: No		
	Woder parameters	weights: No		
	DL algorithm	representation: Yes		Matches
	DE algoritum	•		Watches
		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 hunomarameters	train-test split:		Matches and completes
	Training hyperparameters			
		learning rate: initial between 0.0001-		learning rate: initial between 0.0001-0.1,
		0.1, dropped by 0.001		dropped by 0.001
		#iterations:		#epochs: 2000
		batch size: Size of training data		Train-test split calculated
		#epochs: 2000?		
	Training data	Yes	Input and output are normalized (0-1 and 0-100). Explicitly linked to artifact in paper	
Onomtionalization	Factors and treatments	Partially	Could be several experiments. Not sure if all	Info is in the code
perationalization	ractors and deadnerits	raitially	·	illo is ili tile code
			hyperparams are made explicit Factors (at least): regularization, #hidden layers, learning	
			rate	
	Response variable, elaboration and metric	No	No levels given time (?)	Testing error
Design	Design type	No	une (:)	resung error
Jesigii	Blocking variables	No	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	Partially	#neurons/layer, #epochs, but no value given	reseems ans could be changed in the can
	Measured variables (covariates)	No	mieuroris, iayer, mepocris, bacrio valde giveri	
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
nstrumentation	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		
opulation	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other	
			publications where they are fully explained	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
/alidity evaluation		Partially	Internal, external	
Artifact	Availability	Yes		
	Badge	Yes	Available	

DeepPerf: performance prediction for AP10 4 E1 are a series of experiments. Difficu

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1, RQ4	
	Statistical hypotheses	No		
/ariables selection	Model hyperparameters	Best from previous?		Same as previous
	Model parameters	Best from previous?		
	DL algorithm	Best from previous?		Same as previous
	Training hyperparameters	Best from previous?		Same as previous
Operationalization	Training data Factors and treatments	Yes Model Type, Subject system (?)	Input and output are normalized (0-1 and 0-100) DECART, DeepPerf (DECART is classification trees) Not sure if subject system (apache, x264,8DB-J, LLVM, BDB-C, SQLVIe). Could be blocking variable	Same as previous Same as previous
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time	
Design	Design type	No	Weathenance Error (WILE), daming time	
JC31611	Blocking variables	Yes	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
nstrumentation	Test set	Yes	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		
Consistion			Briefly describes them Deferences are given to the	
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	······································	
	•	Yes	Available	
	Badge	162	Available	

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

Aspect	Element	E3	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ2, RQ4	
	Statistical hypotheses	No		
/ariables selection	Model hyperparameters	Best from previous?		Same as previous
	Model parameter	Part from provious?		
	Model parameters	Best from previous?		
	DL algorithm	Best from previous?		Same as previous
	Training hyperparameters	Best from previous?		Same as previous
	Training Hyperparameters	best non previous:		Jame as previous
	Training data	Yes	Input and output are normalized (0-1 and 0-100)	Same as previous
Onerationalization	Factors and treatments	DNN architecture	SPLConqueror (no DNN), DeepPerf (DECART is	Same as previous
			classification trees)	
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time	
lesign	Design type	No	wear reduce error (wite), during time	
esign	Blocking variables	Yes	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	No	it fold validation (50 times resumpting during set)	rescens ans could be changed in the can
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
nstrumentation	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		
opulation	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other publications where they are fully explained	
Analysis	Descriptive statistics	Yes	Mean and 95% CI	
,	Inferential statistics	No		
/alidity avaluation	Conclusion, internal, construct, external	Partially	Internal, external	
	Availability	Yes	micinal, external	
Artifact	•			
	Badge	Yes	Available	

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

Aspect	Element	E4	Comments	Code
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?		

Best from previous? Model parameters 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

Detection of hidden feature requests from massive chat messages via deep siamese network AP15 $\,$

Paper ID Experiments Comments

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

Aspect	Element	E1	Comments	Code
Experiment type		Optimization+Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#llayers: 2+4(x2)+1+1 #neurons/layer: I1, I2 (512) connections: I1, I2 (feedforward), I3 (input), I4 (convolutional), I5 (BiLSTM), I6 (combination), I7 (similarity) activation functions: I1, I2 softsign params. Initialization: I1, I2 (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	Now fully addressed. No contradictory info e.g. activation functions: l4 relu, l7 sigmoid
	Model parameters	biases: No weights: No	Although I1 and I2 corresponds to a SOTA disentanglement NN, and they are available	
	DL algorithm	representation: Partially	It is not clear if their approach includes	
	or Digoration	model type: Partially loss function: Cross-entropy regularization: Dropout (0.1) and early stopping (after 10 epochs) optimization: No	disentanglement or not. They mention dropout and early stopping as optimization, but it seems to me they are regularization	loss function: Cross-entropy. But also contrastive, MSE and logits appear in frminer model.py optimization: dense sparse Adam in config. ison, Adam in finetune config. ison
				regularization: dropout (0.1) input embeddings & similarity layer
	Training hyperparameters	train-test split: Yes	3-fold intra-project-cross-validation from 3	learning rate: 1e-4
	Training hyperparameters	learning rate: No	projects	#epochs: 10 in config. json, 80 in
		#iterations: No	projecto	finetune config.json, install.MD 100
		batch size: No #epochs: No		batch size: 32
	Technica data	Vac	Fundicials, line of the partitions	No reference in the code to grid search No train d and train t
Operationalization	Training data Factors and treatments	Yes DNN model	Explicitly lined to artifact FRMiner,, p-FRMiner, CNC, FT	No train dand train t
		Yes	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 1.	
Decian	Response variable, elaboration and metric	No No	Precision, recall, F1	
Design	Design type Blocking variables	No No	Project?	
	Held-constant variables Measured variables (covariates) Randomization	No No No	Hoject	
	Task duration	No		
	Procedure	No		
	Number of experimental units	Yes	Cross-validation	
Instrumentation	Test set	Yes	Same as training data	No train d and train t
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Partially	Allennip (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 Inferential statistics	Partially No	Average reported from 3-fold-cross-val	
	Conclusion, internal, construct, external	Partially	External, internal, construct (the authors	
Validity evaluation			define it for RV only)	
Validity evaluation Artifact	Availability	Yes	define it for RV only)	

Detection of hidden feature requests 1 AP15

Paper ID

Experiments Comments

Not sure the extent of the proposed $\boldsymbol{s}_{\boldsymbol{t}}$

Aspect	Element	E2	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ2	
,,	Statistical hypotheses	No	•	
Variables selection	Model hyperparameters	Same as previous?	Same for FRminer and p-Frminer	Same as previous
		•	·	
	Model parameters	Same as previous?		
	DL algorithm	Same as previous?	Same for FRminer and p-Frminer	Same as previous
	DE algoritim	Same as previous:	Same for Exhiniter and p-Emiliner	Same as previous
	Training hyperparameters	Same as previous	Same for FRminer and p-Frminer	Same as previous
	Training hyperparameters	Same as previous	Same for runner and partitioner	Same as previous
	Training data	Same as previous?	Same for FRminer and p-Frminer	Same as previous
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	No	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		
	-	*		
Denulation	Objects (show of the superimental development	Campa de avantiens		
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	<u> </u>	
	Badge	Yes	Available	

Paper ID Detection of hidden feature requests 1 AP15

Experiments Comments Not sure the extent of the proposed $\boldsymbol{s}_{\boldsymbol{t}}$

Aspect	Element	E3	Comments	Code
Experiment type		Optimization+Evaluation		
Hypotheses	Research hypotheses	Yes	RQ3	_
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		Same as previous

Model parameters Same as previous?

Same as previous? DL algorithm Same as previous

3-fold cross-project-cross-validation from Same as previous 3 projects Training hyperparameters Same as previous

Training data	Same as previous?	Same as previous	
Operationalization Eactors and treatments	DNN model	Samo as E1	

	Response variable, elaboration and metric	Yes	Precision, recall, F1	
Design	Design type	No		
	Blocking variables	No	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

On using machine learning to identify knowldge in API reference documentation AP29 $\,$

Paper ID Experiments

Element	E1	Comments	Code
	Comparison	SOTA	
Research hypotheses	Yes	RQ1, RQ2	
Statistical hypotheses	No		
Model hyperparameters	#layers: 5		#layers: 5
	#neurons/layer: 300, ?, 128, 64, 12		#neurons/layer: <u>50</u> , <u>256</u> , 128, 64, 12
	connections: input, LSTM, Dense, Dense, output		connections: input, LSTM, Dense, Dense, output
	activation functions: tahn,, ReLU, ReLU, sigmoid		activation functions: tahn,, ReLU, ReLU, sigmoid
	params. Initialization: No		params. Initialization: Default
Model parameters			
	*		
DL algorithm			representation: Yes
		GloVe	model type: Yes
			loss function: binary cross-entropy
	•		regularization: No
			optimization: dropout (en LSTM= <u>0.2</u>), Adam
Training hyperparameters		_	train-test split: validation=0.2, test=0.1
		as test set	learning rate: 0.001 #iterations: No
			batch size: 32
	#epochs: 100		#epochs: 100
			Fash shaming
Training data	Voc (names montions and and data of the study are	CADO. Recompling is made to improve it	Early stopping NO. Must be requested from authors
ITallillig data		CADO. Resampling is made to improve it	NO. Wrust be requested from authors
Factors and treatments		Two algorithms (k-NN and SV) and RNN with	Implements RNN, k-NN and SV. But in readme, d-NN and SV are not
ractors and deatherto	randary (acamerene evels)		mentioned to be run. Additionally, k-NN code needs CADO. SV needs
			training data not provided
Response variable, elaboration and metric	Partially		·
,			I can only see for RNN F1 and 'accuracy'
		subset accuracy, macroprecision, macrorecall,	
		macroF1, macroAUC	
Design type			
Blocking variables		Glove is trained on 4 corpora for RNN	
Measured variables (covariates)	No		
Randomization	No		
Randomization Task duration	No No		
Task duration	No	10-fold cross-validation using 10% of dataset	10
Task duration Procedure Number of experimental units	No No 10	as test set	
Task duration Procedure Number of experimental units Test set	No No 10 Partially	_	10 Available
Task duration Procedure Number of experimental units	No No 10	as test set	
Task duration Procedure Number of experimental units Test set	No No 10 Partially	as test set	
Task duration Procedure Number of experimental units Test set Measuring instruments	No No 10 Partially Deduced	as test set	
Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure	No No 10 Partially Deduced Deduced	as test set New Python dataset (not very well explained)	
Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No 10 Partially Deduced Deduced Partially	as test set New Python dataset (not very well explained) GloVe for embeddings, trained on 4 corpora	
Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets)	No No 10 Partially Deduced Deduced Partially Partially Partially	as test set New Python dataset (not very well explained) GloVe for embeddings, trained on 4 corpora New Python dataset (not very well explained)	
	Research hypotheses Statistical hypotheses Model hyperparameters Model parameters DL algorithm Training hyperparameters Training data Factors and treatments Response variable, elaboration and metric Design type Blocking variables Held-constant variables Measured variables (covariates)	Research hypotheses Statistical hypotheses No Model hyperparameters Model hyperparameters Model hyperparameters Model hyperparameters Model parameters Model parameters	Research hypotheses Yes Research hypotheses Yes Research hypotheses No Model hyperparameters Maly perparameters Model hyperparameters Model parameters Input word embedding vectors trained using GioVe Model parameters Model para

MTFuzz: Fuzzing with a multi-task neural network AP36 4

Paper ID Experiments

Aspect	Element	E1	Comments	Code
Experiment type		Comparison		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters		Each task has the same weight	#layers: 7 + 3(with 3 paralells)
		#neurons/layer: L1(?),		#neurons/layer: L1(2048), L2(2048), L3(1024),
		L2(2048), L3(1024), L4(512)		L4(512), for decoder?: 8,32 (commented),1
		connections: 3 encoder, 3 (x3)		connections: 3 encoder, 3 (x3) decoder
		decoder		activation functions: ReLu for hidden, sigmoid
		activation functions: ReLu for		output
		hidden, sigmoid output		params. Initialization: random
		params. Initialization:		
	Model parameters	biases: No		Model is saved and reloaded, but not stored.
		weights: No		For avoiding crashes
	DL algorithm	representation: Yes	Loss function: MSE for edge coverage,	representation: Yes
		model type: Yes	adaptive loss for edge and context-sensitive	model type: Yes
		loss function: multi-task	edge	loss function: multi-task
		regularization: No		regularization: No
		optimization: Adam		optimization: Adam
	Training hyperparameters	train-test split: No	750 input samples for re-training	train-test split: Yes
		learning rate: 0.001		learning rate: 0.001, also 0.1 for decoder?
		#iterations: No		#iterations: 50 (with 100 epochs)
		batch size: No		batch size: Yes (with 300 epochs)
		#epochs: 100		#epochs: 100. But also allows 300
	Training data	Yes		Names do not match the ones in the paper
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	10	
	Blocking variables Held-constant variables	Program No	10 programs	
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	Yes	24 hours for real-world, 5 hours for synthetic	
	Task dardton	163	bugs	
	Procedure	No	bugs	
	Number of experimental units	Yes	5 repetitions to cover fuzzer variability	
Instrumentation	Test set	Yes	2 datasets, one for real bugs, other for synthetic	Names do not match the ones in the paper
	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		
		Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

MTFuzz: Fuzzing with a multi-task net AP36 4

Paper ID

Aspect	Element	E2	Comments	Code
Experiment type		Optimization	With some/without auxiliary tasks	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	Woder parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	DE digonami	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		Same as previous
Operationalization	Factors and treatments	Partially	Removing some of the decoders. All configs. Use	
Operationalization	ractors and deatherts	Tardany	same hyperparams, etc.	-
			same ny perparams, 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	Names do not match the ones in the pape
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu	18.04,
			Intel Xeon E5-2623, NVIDIA GTX 1080Ti GP	U.
			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		

MTFuzz: Fuzzing with a multi-task net AP36 4

Paper ID Experiments

Aspect	Element	E3	Comments	Code
Experiment type		Optimization	Adaptive loss	
Hypotheses	Research hypotheses	Yes	RQ3	·
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
Variables selection	Model hyperparameters	Same as previous		Same as previous
Variables selection	Model hyperparameters	Same as previous		Same as previous

	Model parameters	Same as previous		Same as previous
	Woder parameters	Same as previous		Sume as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		Same as previous
perationalization	Factors and treatments	Yes	Adaptive loss	
	Response variable, elaboration and metric	No	Recall, F1	Completes
esign	Design type	No		
	Blocking variables Held-constant variables	Program No		
		No		
	Measured variables (covariates)			
	Randomization Task duration	No	Non-constitution delications	
	Task duration	No	Not specified this time	
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
strumentation	Test set	Yes	Real bugs only	Names do not match the ones in the paper
	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	
opulation	Objects (chars. of the experimental datasets)	Partially		
nalysis	Descriptive statistics	No		
,	Inferential statistics	No		

			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	

MTFuzz: Fuzzing with a multi-task net AP36

Paper ID

Aspect	Element	E4	Comments	Code
Experiment type	<u> </u>	Generalization		
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	-			
	Training hyperparameters	Same as previous		Same as previous
	0 //- /			
	Training data	Same as previous		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	Edge to verage	
	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	Names do not match the ones in the paper
	Measuring instruments	Deduced	•	
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0.	
			Ubuntu18.04, Intel Xeon E5-2623,	
			NVIDIA GTX 1080Ti GPU.	
			For data collection, single core machine	
			for an hour	
Population	Objects (chars. of the experimental datasets)	Partially		
Analysis	Descriptive statistics	No		
,	Inferential statistics	No		
Validity evaluation		Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

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

ID AP39 Experiments 12

Comments E9-11 appear in supplementary material only. E12 is referenced in supplementary material

Aspect	Element	E1	Comments	Code
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	Architecture is described at a high level. For details points to supplementary material	Complete in artifact. But possible incosistencies between supplementary material and code
	Model parameters	params. Initialization: No biases: No	Does not explicitly mention parameters	Pre-trained available in artifact, but broker
	DL algorithm	weights: No representation: Yes model type: Yes loss function: Yes regularization: No optimization: No	Loss: MSE + BCE. Points to supplementary material	link In code
	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	<u>In code</u>
	Training data	Yes	Details in supplementary material	There are 3 sets available in dropbox: raw data, preprocessed pre-training and training (fine-tuning). The links of the last 2 are broken. The organization of the first one does not seem to match the paper, and the datasets are not the same as the ones described in the supplementary material. Some are missing, and the others seem to have several versions (not the latests as stated in the paper)
Operationalization	Factors and treatments	Model type (STATEFORMER)	STATEFORMER performance is evaluated	Stated in the paper y
	Response variable, elaboration and metric	Yes	Precision, Recall , F1	
Design	Design type	No		
	Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units	No No No No No No	Architecture/optimization/obfuscation Seems 1 run	
Instrumentation	Test set	Yes	Details in supplementary material	There are 2 sets available in dropbox: raw
Instrumentation				data and preprocessed pre-training. The link of the later is broken. The organization of the former does not seem to match the paper, and the datasets are not the same as the ones described in the supplementar material. Some are missing, and the other seem to have several versions (not the latests as stated in the paper)
	Measuring instruments	No		
	Measuring instruments Measurement procedure	No No		
	•		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	Measurement procedure	No	Linux server w Ubuntu 18.04 Intel Xeon 4212 2.2.0GHz 48 virtual cores 188GB RAM 4 Nvidia RTX 2080-Ti GPUs	<u>Yes</u>
	Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics	No Yes Yes No	Linux server w Ubuntu 18.04 Intel Xeon 4212 2.2.0GHz 48 virtual cores 188GB RAM 4 Nvidia RTX 2080-Ti GPUs pyeiftools, Ghidra	<u>Yes</u>
Analysis	Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No Yes Yes No No	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 Details in supplementary material	<u>Yes</u>
Population Analysis Validity evaluation Artifact	Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No Yes Yes No	Linux server w Ubuntu 18.04 Intel Xeon 4212 2.2.0GHz 48 virtual cores 188GB RAM 4 Nvidia RTX 2080-Ti GPUs pyeiftools, Ghidra	<u>Yes</u>

Paper ID STATEFORMER: Fine-grained type rec

AP39 12 Experiments

Comments E9-11 appear in supplementary mater

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	_
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

Operationalization	Factors and treatments	Model type (STATEFORMER,	For EKLAVIA, numbers reported in paper are	!
		EKLAVIA)	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	Same as previous

Measuring instruments No Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	EKLAVIA projects. Supplementary material	Same as previous
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 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E3	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

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	Same as previous

Measuring instruments No
Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E4	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

Operationalization	Factors and treatments	Model type (STATEFORMER,	Typeminer is not open-source. Authors
		Typeminer)	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 Same as previous

Measuring instruments No
Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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

ID AP39 Experiments 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E5	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

0	F. d dt dt	AA-JUL - (CTATEFORMER D. L.	Children to a constitution of the Children	
Operationalization	Factors and treatments	,, ,	Ghidra is commercial tool (not DNN)	
		Ghidra)		
	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	Same as previous

 Measuring instruments
 No

 Measurement procedure
 No

 Technological infrastructure
 Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E6	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ4	_
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

Operationalization	Factors and treatments	Use of pre-training, masking	Not sure the value of the other once one of them is fixed	Artifact does not seem to contain ablation studies (only the final model)
	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	
nstrumentation	Test set	No		Same as previous

Measuring instruments No
Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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 ID STATEFORMER: Fine-grained type rec

AP39 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E7	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
	×	• • • • • • • • • • • • • • • • • • • •		

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		Same as previous

Measuring instruments No Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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 ID STATEFORMER: Fine-grained type rec

AP39 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E8	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous

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		Same as previous

Measuring instruments No Measurement procedure No

Technological infrastructure Same as previous

Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
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

Paper STATE ID AP39 Experiments 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E9-supplementary	Comments
Experiment type		Assessment (different types predicted)	
	Research hypotheses	No	
	Statistical hypotheses	No	
	Model hyperparameters		Same as previous
variables selection	inoce hyperparameters		June 13 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 (type)	Assesses STATEFORMER on predicting
ореги полите и полите	Totals and declinents	ics (type)	different types
	Response variable, elaboration and metric	Partially	Accuracy, not defined
	Design type	No	
-	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure		
	Number of experimental units	No No	Seems 1 run
	Test set	Yes	Same as previous
	Measuring instruments Measurement procedure Technological infrastructure	No No Same as previous	Deduced from paper.
Population	Objects (chars. of the experimental datasets)	Yes	
	Descriptive statistics	Partially	Average only
	Inferential statistics	No	
	Conclusion, internal, construct, external	No	
Validity evaluation	Conclusion, internal, construct, external Availability	Yes	

Paper STATEFORMER: Fine-grained type rec

Paper STATE ID AP39 Experiments 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E10-supplementary	Comments
Experiment type	Lienten	Generalization (cross-project)	Comments
Hypotheses	Research hypotheses	No	
Trypodieses	Statistical hypotheses	No	
Variables selection	Model hyperparameters	110	Same as previous
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	Test set	
	Response variable, elaboration and metric	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	Yes	Same as previous
	Measuring instruments Measurement procedure	No No	
	Technological infrastructure	Same as previous	Deduced from paper
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	Yes	Available, reusable
	07		

Paper ID STATEFORMER: Fine-grained type rec

AP39 12 Experiments

Comments E9-11 appear in supplementary mater

Aspect	Element	E11-supplementary	Comments
Experiment type		Comparison	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
variables selection	жее пуревание сез		<u>Jame do prestous</u>
	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 type (STATEFORMER, Debin)	
	Response variable, elaboration and metric	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	Obfuscated ones	Same as previous
	Measuring instruments	No	
	-		
	Measurement procedure Technological infrastructure	No Same as previous	
	recimologica inirastructure	Same as previous	
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	Yes	Available, reusable

Paper ID STATEFORMER: Fine-grained type rec

AP39 Experiments 12

Comments E9-11 appear in supplementary mater

Aspect	Element	E12-supplementary	Comments
Experiment type		Optimization	
Hypotheses	Research hypotheses	No	
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
variables selection	мочет пуреграванете:		Jane as previous
	Model parameters		Same as previous
	DL algorithm		Same as previous
	Training hyperparameters		Same as previous
	Training data		Same as previous
0	e	N	
Operationalization	Factors and treatments	Numerical values embedding, number of layers, layers dimensions,	
	Response variable, elaboration and metric	Training loss	
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	Same as previous
	Advantage of the Control of the Cont		
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	No	
•	Inferential statistics	No	
Validity evaluation		No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

A syntax-guided edit decoder for neural program repair AP40 5

Paper ID Experiments Comments

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

A syntax-guided edit decoder for neural prograr AP40 5

Paper ID Experiments Comments

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		Same as previous
Operationalization	Factors and treatments	SequenceR, CODIT, DLFix,		Possible inconsistencies due to DL algorithm
		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		
	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper ID A syntax-guided edit decoder for neural prograr AP40 5

Experiments 5

Aspect	Element	E3	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		It is a factor now, but the different datasets do not appear in paper (thus, artifact does not improve)
Operationalization	Factors and treatments	Removing: modify, subtreecopy, insert, placeholder. With eveverything	But testsets are not expected in the code. It cannot be FA	Variation of training datasets do not appear in paper (thus, artifact does not improve paper)
	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		
	buuge	Available		

A syntax-guided edit decoder for neural prograr AP40 5

Paper ID Experiments

Aspect	Element	E4	Comments	Code
Experiment type		Generalization+Evaluation	Tested in a diferent dataset	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as E1-E2
	Training data	Same as E1-E2		Same as E1-E2
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 v2.0 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		
	•			

A syntax-guided edit decoder for neural prograr AP40 5 Paper ID

Aspect	Element	E5	Comments	Code
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		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		Same as previous
	Training hyperparameters	Same as previous		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	Subsets not included in code (as expected
	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		

Paper

Lightweight global and local contexts guided method name recommendation with prior knowledge

AP41 7 ID Ex

Experiments	
Comments	

Aspect	Element	E1	Comments	Code	
Experiment type		Optimization	Hidden		
Hypotheses	Research hypotheses	No			
rrypotrieses	Statistical hypotheses	No			
Variables selection	Model hyperparameters	Same as next			
variables selection	woder ny perparameters	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			
Instrumentation	Number of experimental units Test set	No No			
mstrumentation	resuset	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			
	· · ·				
Altidet	Badge	Yes	Available		

Paper Lightweight global and local contexts guided mo

ID AP41
Experiments 7
Comments

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Compares against SOTA	
			200	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: Yes		Code provided
		#neurons/layer: No		
		connections: No activation functions: No		
		params. Initialization: No		
		pararris. midalization. No		
	Model parameters	biases: No		
		weights: No		
	DL algorithm	representation: Yes	"Due to page limit, we only briefly	optimization: Adagrad
		model type: Yes	introduce this model in the paper, and	regularization: dropout
		loss function: Yes	more details could be referred to the	
		regularization: No	existing work [57]"	
		optimization: No		
			Loss function:negative log likelihood of the oracle word for that step	
	Training hyperparameters	train-test split: Yes		learning rate: 0.15
		learning rate: No		iterations: 500,000
		#iterations: No		batch_size: 120
		batch size: No		epochs: inconsistent, 2 values used, both
		#epochs: No		in comments (10 and 5)
	Training data	Yes	References known datasets (Java-	
			small, Java-med, Java-large, Mnire) publicly available	
Operationalization	Factors and treatments	Model type (10	For Mnire, they use results reported in	
·		approaches vs Cognac)	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	ridenst of tiffeats	
Artifact	Badge	Yes	Available	
	Dauge	1 53	Available	

Paper Lightweight global and local contexts guided m

AP41 7 ID

Experiments	
Comments	

Aspect	Element	E3	Comments	Code
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		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	Reference to known dataset (Liu et al),	
		103	publicly available	
			. ,	
Operationalization	Factors and treatments	Model type (Liu et al, Mnire,	Do not explain any of the others	
		Cognac)		
		Class (consistent, inconsistent)		
	Response variable, elaboration and metric	Yes	Precision, Recall, F-score	
Design	Design type	No	riedsion, recail, i -score	
Design	Blocking variables	No	Could class be a blocking variable?	
	Siconing variables		codia diasa se a siocking variable.	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
		No		
	Task duration	110		
	Procedure	No		
	Procedure Number of experimental units	No No	Seems 1 run	
Instrumentation	Procedure	No	Seems 1 run Same as training set	
Instrumentation	Procedure Number of experimental units Test set	No No Yes		
Instrumentation	Procedure Number of experimental units Test set Measuring instruments	No No Yes		
Instrumentation	Procedure Number of experimental units Test set Measuring instruments Measurement procedure	No No Yes No No		
	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No Yes No No No		
Population	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets)	No No Yes No No No		
	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics	No No Yes No No No No	Same as training set	
Population	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets)	No No Yes No No No		
Population	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No No Yes No No No No	Same as training set	
Population Analysis	Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No No Yes No No No No No No No	Same as training set Conclusions based on 1 run. At a guess	

Paper

Lightweight global and local contexts guided me

Measuring instruments

Descriptive statistics Inferential statistics

Validity evaluation Conclusion, internal, construct, external

Availability

Badge

Population

Analysis

Artifact

Measurement procedure

Technological infrastructure

Objects (chars. of the experimental datasets) No

No

No

No

No

Partial

Yes

Yes

ID AP41 Experiments 7 Comments

Aspect	Element	E4	Comments	Code
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		Same as previous

Model parameters Same as previous DL algorithm Same as previous Same as previous Training hyperparameters Same as previous Same as previous Training data Yes Same as E2 Operationalization Factors and treatments No further details are given Model (no caller info, no callee info, no prior knowledge, Cognac) 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.

Conclusions based on 1 run. At a guess

Flat list of threats

Available

Paper Lightweight global and local contexts guided mo

ID AP41 Experiments 7

Comments

Aspect	Element	E5	Comments	Code
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		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	Same as E3	
Operationalization	Factors and treatments	Model (no caller info, no callee info,	No further details are given	
		no prior knowledge, Cognac)		
		Class (consistent/inconsistent)		
	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		
			C1	
	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 m

ID AP41 Experiments 7

Comments

Aspect	Element	E6	Comments	Code
Experiment type		Optimization	use of caller/calle info.	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	,, ,	·		
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	References known dataset (Mnire), publicly available	
Operationalization	Factors and treatments	Model type (seq2seq mo	de No details are given about seq2s	seq
	Response variable, elaboration and metric	Yes	F-score	
Design	Response variable, elaboration and metric Design type	Yes	F-score	
Design	Response variable, elaboration and metric Design type Blocking variables		F-score	
Design	Design type Blocking variables	No No	F-score	
Design	Design type Blocking variables Held-constant variables	No	F-score	
Design	Design type Blocking variables	No No	F-score	
Design	Design type Blocking variables Held-constant variables Measured variables (covariates)	No No No	F-score	
Design	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization	No No No No	F-score	
Design	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure	No No No No No No	F-score Seems 1 run	
Design Instrumentation	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration	No No No No No No		
	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units	No No No No No No No	Seems 1 run	
	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set	No No No No No No No No No Yes	Seems 1 run	
	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments	No No No No No No No Vo Yes	Seems 1 run	
Instrumentation	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure	No No No No No No No Vo Yes	Seems 1 run	
Instrumentation	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure	No No No No No No No Yes	Seems 1 run	
Instrumentation	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets)	No N	Seems 1 run	
Instrumentation Population Analysis	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No N	Seems 1 run Same as training set. Conclusions based on 1 run. At a	
Instrumentation	Design type Blocking variables Held-constant variables Measured variables (covariates) Randomization Task duration Procedure Number of experimental units Test set Measuring instruments Measurement procedure Technological infrastructure Objects (chars. of the experimental datasets) Descriptive statistics Inferential statistics	No N	Seems 1 run Same as training set. Conclusions based on 1 run. At a guess	

Paper

Comments

Lightweight global and local contexts guided me

Measuring instruments
Measurement procedure

Descriptive statistics Inferential statistics

Validity evaluation Conclusion, internal, construct, external

Availability

Badge

Population

Analysis

Artifact

Technological infrastructure

Objects (chars. of the experimental datasets) No

ID AP41 Experiments 7

Aspect	Element	E7	Comments	Code
Experiment type		Optimization	Similar to E1, but with all tokens	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		<u>Same as previous</u>
	Training data	Yes	References known datasets(Java-smal Java-med, Java-large, Mnire), publicly available	l,
Operationalization	Factors and treatments	Yes	Tokens (all vs. 10)	
	Response variable, elaboration and metric	Yes	F-score	
Design	Design type Blocking variables	No 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	

No

No

No

Partial

Yes

Yes

test with the same test.

Flat list of threats

Available

Conclusions based on 1 run. At a guess