**Table 4: Characterization of (44) experiments that earned ACM Artifact Badges**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Improvements** | | | | **Constant** | | |
| **Step** | **Aspect** | **PA→FA** | **M→PA** | **M→FA** | **PA→PA** | **M** | **PA** | **FA** |
| S1 | Research hypotheses | 0% | 0% | 0% | 0% | 18% | 0% | 82% |
| S2 | Model hyperparameters | 68% | 0% | 7% | 0% | 0% | 5% | 20% |
|  | Model parameters | 0% | 0% | 9% | 0% | 82% | 0% | 9% |
|  | DL algorithm | 39% | 0% | 7% | 0% | 0% | 13% | 41% |
|  | Training hyperparameters | 59% | 0% | 7% | 0% | 0% | 14% | 20% |
|  | Training data | 7% | 0% | 7% | 0% | 2% | 4% | 80% |
| S3 | Factors and treatments | 0% | 2% | 5% | 52% | 0% | 30% | 11% |
|  | Response variables | 2% | 3% | 9% | 0% | 2% | 0% | 84% |
| S4 | Choice of design | 0% | 0% | 0% | 0% | 39% | 59% | 2% |
|  | Instrumentation | 2% | 7% | 0% | 5% | 2% | 75% | 9% |
| S5 | Test set chars. | 0% | 0% | 7% | 0% | 43% | 30% | 20% |
| S6 | Descriptive statistics | 0% | 2% | 0% | 0% | 64% | 23% | 11% |
|  | Inferential statistics | 0% | 0% | 0% | 0% | 95% | 0% | 5% |
| S7 | Validity threats | 0% | 0% | 0% | 0% | 27% | 71% | 2% |

Characterization of (44) experiments that earned ACM artifact badges. Cells with a single label mean the artifact does not improve the information provided in the paper. Cells with an arrow mean that the information provided in the artifact improves the information provided in the paper. Light gray background represents an inconsistency between the content of the paper and the artifact. Dark gray background represents inconsistencies within the elements in the artifact. **This table does not appear in the paper.**

Imagen que contiene texto, juego

Descripción generada automáticamente

**AP5:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Discrepancy between paper and file “model.h5” provided | Same as previous | Same as previous | Same as previous |
| Model parameters | File .h5 provided | Same as previous | Same as previous | Same as previous |
| DL algorithm | Training code not provided | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Training code not provided | Same as previous | Same as previous | Same as previous |
| Training data | Provided (dropbox link). Matches paper | Same as previous | Same as previous | Same as previous |
| Operationalization | Factors and treatments | Discrepant model hyperparameters. Training code not provided | Same as previous for NL2Type. DeepTyper DNN is taken from the authors and further trained, but not included in artifact | Same as E1 | - |
| Response variables | - | - | - | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | Test set provided (dropbox link) | Same as previous | - | - |
| Objects selection | Test set characteristics | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

**AP8:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Inconsistencies with code | Same as previous | Same as previous | Same as previous |
| Model parameters | Missing. Inconsistency with paper (says it is provided) | Same as previous | Same as previous | Same as previous |
| DL algorithm | Inconsistencies with code | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Inconsistencies with code | Same as previous | Same as previous | Same as previous |
| Training data | - | - | - | - |
| Operationalization | Factors and treatments | Improved, but inconsistencies might affect | Improved, but inconsistencies might affect | Improved, but inconsistencies might affect | Improved, but inconsistencies might affect |
| Response variables | - | - | - | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | - | - | - | - |
| Objects selection | Test set characteristics | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

**AP10:**

E4 is missing from the code

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Matches and completes | Matches and completes | Matches and completes | - |
| Model parameters | - | - | - | - |
| DL algorithm | Matches and completes | Matches and completes | Matches and completes | - |
| Training hyperparameters | Matches and completes | Matches and completes | Matches and completes | - |
| Training data | Matches | Matches | Matches | - |
| Operationalization | Factors and treatments | Completes | Same as previous | Same as previous | - |
| Response variables | Testing error (completes) | - | - | - |
| Design | Choice of design | The 30 times can be changed when invoking the function (not really inconsistency) | The 30 times can be changed when invoking the function (not really inconsistency) | The 30 times can be changed when invoking the function (not really inconsistency) | - |
| Instrumentation | Matches | Matches | Matches | - |
| Objects selection | Test set characteristics | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

**AP15:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** |
| Hypotheses formulation | Research hypotheses | - | - | - |
| Variables identification | Model hyperparameters | Now fully addressed. No contradictory info. | Now fully addressed. No contradictory info. | Now fully addressed. No contradictory info. |
| Model parameters | - | - | - |
| DL algorithm | Contradictory info (optimization and loss function) | Contradictory info (optimization and loss function) | Contradictory info (optimization and loss function) |
| Training hyperparameters | Contradictory info (epochs) in different files. No reference to grid search | Contradictory info (epochs) in different files. No reference to grid search | Contradictory info (epochs) in different files. No reference to grid search |
| Training data | No train\_d and train\_t (splitted data). Inconsistency with paper | No train\_d and train\_t (splitted data). Inconsistency with paper | No train\_d and train\_t (splitted data). Inconsistency with paper |
| Operationalization | Factors and treatments | Issues due to previous contradictions | Issues due to previous contradictions | Issues due to previous contradictions |
| Response variables | - | - | - |
| Design | Choice of design | - | - | - |
| Instrumentation | - | - | - |
| Objects selection | Test set characteristics | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - |
| Inferential statistics | - | - | - |
| Validity evaluation | Validity threats | - | - | - |

**AP29:**

|  |  |  |
| --- | --- | --- |
|  |  | **E1** |
| Hypotheses formulation | Research hypotheses | - |
| Variables identification | Model hyperparameters | Discrepancy in the number of neurons per layer.  Now complete |
| Model parameters | - |
| DL algorithm | Discrepancy with loss function.  Now complete. |
| Training hyperparameters | Now complete |
| Training data | Not provided. Must be requested from authors. Paper mentions that they share code and data |
| Operationalization | Factors and treatments | Not all treatments are included. Discrepancy |
| Response variables | Not all of them. Discrepancy |
| Design | Choice of design | - |
| Instrumentation | - |
| Objects selection | Test set characteristics | - |
| Analysis & interpretation | Descriptive statistics | - |
| Inferential statistics | - |
| Validity evaluation | Validity threats | - |

**AP36:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Now complete, but there are commented lines in the decoder | Same as previous | Same as previous | Same as previous |
| Model parameters | Model is saved and reloaded to avoid crashes, but not stored | Same as previous | Same as previous | Same as previous |
| DL algorithm | Now complete | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Now complete. But there are discrepancies | Same as previous | Same as previous | Same as previous |
| Training data | Names do not match the ones in the paper | Same as previous | Same as previous | Same as previous |
| Operationalization | Factors and treatments | Inconsistencies due to variables above | Same as previous | Same as previous | Same as previous |
| Response variables | - | - | Completes | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | Names of datasets do not match the ones in the paper | Same as previous | Same as previous | Same as previous |
| Objects selection | Test set characteristics | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

**AP39:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Possible inconsistencies between supplementary material and code | Same as previous | Same as previous | Same as previous |
| Model parameters | Pre-trained available in paper (but link is broken) | Same as previous | Same as previous | Same as previous |
| DL algorithm | Yes | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Yes | Same as previous | Same as previous | Same as previous |
| Training data | Match neither paper nor supplementary material. | Same as previous | Same as previous | Same as previous |
| Operationalization | Factors and treatments | Possible inconsistencies due to model hyperparam | Same as previous | Same as previous | Same as previous |
| Response variables | - | - | - | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | Datasets match neither paper nor supplementary material | Same as previous | Same as previous | Same as previous |
| Objects selection | Test set characteristics | Ok | Ok | Ok | Ok |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E5** | **E6** | **E7** | **E8** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Same as previous | Same as previous | Same as previous | Same as previous |
| Model parameters | Same as previous | Same as previous | Same as previous | Same as previous |
| DL algorithm | Same as previous | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Same as previous | Same as previous | Same as previous | Same as previous |
| Training data | Same as previous | Same as previous | Same as previous | Same as previous |
| Operationalization | Factors and treatments | Same as previous. Artifact does not seem to contain ablation study | Same as previous | Same as previous | Same as previous |
| Response variables | - | - | - | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | Same as previous | Same as previous | Same as previous | Same as previous |
| Objects selection | Test set characteristics | Ok | Ok | Ok | Ok |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

These are the ones that appear in the supplementary material. The last one (E13) is the only one specifically referenced in the paper.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E9** | **E10** | **E11** | **E12** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | Same as previous | Same as previous | Same as previous | Same as previous |
| Model parameters | Same as previous | Same as previous | Same as previous | Same as previous |
| DL algorithm | Same as previous | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Same as previous | Same as previous | Same as previous | Same as previous |
| Training data | Same as previous | Same as previous | Same as previous | Same as previous |
| Operationalization | Factors and treatments | Predicted type | Test set | Model type (Stateformer, Debin) | Numerical values embedding, number of layers, layers dimensions, |
| Response variables | Accuracy (just mentioned) | F1 | F1 | Training loss |
| Design | Choice of design | No | No | No | No |
| Instrumentation | Same as others | Same as others | Same as others | Same as others |
| Objects selection | Test set characteristics | Yes | Yes | Yes | Yes |
| Analysis & interpretation | Descriptive statistics | Average only | No | No | No |
| Inferential statistics | No | No | No | No |
| Validity evaluation | Validity threats | No | No | No | No |

**AP40:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** | **E5** |
| Hypotheses formulation | Research hypotheses | - | - | - | - | - |
| Variables identification | Model hyperparameters | Now appear. Do not contradict paper because in paper they are not described | Same as previous | Same as previous | Same as previous | Same as previous |
| Model parameters | - | - | - | - | - |
| DL algorithm | AdamW also appears in code | Same as previous | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Now complete and do not contradict paper | Same as previous | Same as previous | Same as previous | Same as previous |
| Training data | In paper it is not specifically linked, but appears in RP | Same as previous | - | Same as E1-E2 | - |
| Operationalization | Factors and treatments | Possible inconsistencies due to DL algorithm | Same as previous | Treatments are different subsets of training set. They do not appear in artifact. Does not improve | Same as E1-E2 | Treatments are different subsets of training set. They do not appear in artifact. Does not improve |
| Response variables | - | - | - | - | - |
| Design | Choice of design | - | - | - | - | - |
| Instrumentation | - | - | - | - | - |
| Objects selection | Test set characteristics | - | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - | - |
| Inferential statistics | - | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - | - |

**AP41:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **E1** | **E2** | **E3** | **E4** |
| Hypotheses formulation | Research hypotheses | - | - | - | - |
| Variables identification | Model hyperparameters | - | Yes. There is a configuration file (data\_util/config.py) | Same as previous | Same as previous |
| Model parameters | - | - | - | - |
| DL algorithm | - | Configuration file | Same as previous | Same as previous |
| Training hyperparameters | - | Configuration file. There are two (commented) values for epochs. Inconsistency | Same as previous | Same as previous |
| Training data | - | Linked | Linked | Linked |
| Operationalization | Factors and treatments | - | Possible inconsistencies due to training hyperparameters | Same as previous | Same as previous |
| Response variables | - | - | - | - |
| Design | Choice of design | - | - | - | - |
| Instrumentation | - | - | - | - |
| Objects selection | Test set characteristics | - | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - | - |
| Inferential statistics | - | - | - | - |
| Validity evaluation | Validity threats | - | - | - | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **E5** | **E6** | **E7** |
| Hypotheses formulation | Research hypotheses | - | - | - |
| Variables identification | Model hyperparameters | Same as previous | Same as previous | Same as previous |
| Model parameters | - | - | - |
| DL algorithm | Same as previous | Same as previous | Same as previous |
| Training hyperparameters | Same as previous | Same as previous | Same as previous |
| Training data | Linked | Linked | Linked |
| Operationalization | Factors and treatments | Same as previous | Same as previous | - |
| Response variables | - | - | - |
| Design | Choice of design | - | - | - |
| Instrumentation | - | - | - |
| Objects selection | Test set characteristics | - | - | - |
| Analysis & interpretation | Descriptive statistics | - | - | - |
| Inferential statistics | - | - | - |
| Validity evaluation | Validity threats | - | - | - |