Asghar, A., Noetel, M., Schellekens, M., Parker, P., Antczak, D., Beauchamp, M., Dicke, T., Diezmann, C., Maeder, A., Ntoumanis, N., Yeung, A., Lonsdale, C. (2021). ASystematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity. *Psychosocial Intervention.* <https://doi.org/10.5093/pi2021a4>

**Supplementary File 5.** *Predictive Performance of each Method*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Study | Predictive accuracy measure | Value | Codes being predicted | Interpretation | Size of dataset |
| Latent Dirichlet Allocation | Atkins et al., 2014 | Area Under the Curve | 0.62-0.81 | 10 | Better than Chance = All the codes Poor = 3 codes Acceptable =: 5 codes Excellent = 2 codes Outstanding = 0 codes | 1,004,924 words |
| Latent Dirichlet Allocation | Atkins et al., 2014 | Intra-Class Correlation | Excellent = 7 codes Good = 1 code Fair = 0 codes Poor = 2 codes | 10 | Excellent =7 codes Good =1 code Fair = 0 codes Poor = 2 codes | 1,004,924 words |
| Latent Dirichlet Allocation | Atkins et al., 2014 | Kappa | Almost perfect agreement = 1 code Substantial = 5 codes moderate or less = 4 codes | 10 | Almost perfect agreement = 1 code Substantial = 5 codes moderate or less = 4 codes | 1,004,924 words |
| Maximum Entropy Markov Model | Can et al., 2012 | F1-score | 0.81 | 10 | -- | sessions 57 |
| Hidden Markov Model | Can et al., 2012 | F1-score | 0.71 | 10 | -- | sessions 57 |
| Conditional Random Field | Can et al., 2015 | F1-score | MISC28-code = 0.75 | 19 | -- | 1,736,000 words |
| Maximum Entropy Markov Model | Can et al., 2016 | F1-score | 0.81 | 3 | -- | sessions 57 |
| Naive Bayes | Carcone et al., 2019 | Kappa | 0.497 | 41 | Moderate | utterances 11,353 |
| Naive Bayes | Carcone et al., 2019 | F1-score | 0.55 | 41 | -- | utterances 11,353 |
| Naive Bayes-Multinomial | Carcone et al., 2019 | Kappa | 0.62 | 41 | Substantial | utterances 11,353 |
| Naive Bayes-Multinomial | Carcone et al., 2019 | F1-score | 0.64 | 41 | -- | utterances 11,353 |
| J48 | Carcone et al., 2019 | Kappa | 0.54 | 41 | Moderate | utterances 11,353 |
| J48 | Carcone et al., 2019 | F1-score | 0.58 | 41 | -- | utterances 11,353 |
| AdaBoost | Carcone et al., 2019 | Kappa | 0.57 | 41 | Moderate | utterances 11,353 |
| AdaBoost | Carcone et al., 2019 | F1-score | 0.61 | 41 | -- | utterances 11,353 |
| Random Forest Model | Carcone et al., 2019 | Kappa | 0.62 | 41 | Substantial | utterances 11,353 |
| Random Forest Model | Carcone et al., 2019 | F1-score | 0.62 | 41 | -- | utterances 11,353 |
| DiscLDA | Carcone et al., 2019 | Kappa | 0.39 | 41 | Fair | utterances 11,353 |
| DiscLDA | Carcone et al., 2019 | F1-score | 0.43 | 41 | -- | utterances 11,353 |
| Conditional Random Field | Carcone et al., 2019 | Kappa | 0.51 | 41 | Moderate | utterances 11,353 |
| Conditional Random Field | Carcone et al., 2019 | F1-score | 0.67 | 41 | -- | utterances 11,353 |
| Support Vector Machine | Carcone et al., 2019 | Kappa | 0.663 | 41 | Substantial | utterances 11,353 |
| Support Vector Machine | Carcone et al., 2019 | F1-score | 0.68 | 41 | -- | utterances 11,353 |
| Static Behavior Model | Chakravarthula et al., 2015 | Accuracy | 0.81 | 2 | -- | sessions 200 |
| Activation-based Dynamic Behavior Model | Chakravarthula et al., 2015 | Accuracy | 0.755 | 2 | -- | sessions 200 |
| Likelihood-based Dynamic Behavior Model | Chakravarthula et al., 2015 | Accuracy | 0.755 | 2 | -- | sessions 200 |
| Fidelity Automatic Rater | Gallo et al., 2015 | Pearson Correlation Coefficient | 0.32-0.35 | 3 | Weak | 86,000 words |
| Labelled Latent Dirichlet Allocation | Gaut et al., 2017 | Area Under the Curve | 0.789 | 41 | Acceptable | 8,000,000 words |
| Lasso Logistic Regression | Gaut et al., 2017 | Area Under the Curve | 0.7 | 41 | Acceptable | 8,000,000 words |
| Deep Neural Networks | Gibson et al., 2016 | F1-score | MISC-8 code = 0.643 MISC-28 code = 0.258 | 19 | -- | sessions 348 |
| Recurrent Neural Networks with attention-based LSTM | Gibson et al., 2017 | F1-score | 0.637 | 8 | -- | 1,659,000 words |
| Feed-Forward Neural Network | Gibson et al., 2017 | F1-score | 0.58 | 8 | -- | 1,659,000 words |
| Maximum Entropy Markov Model | Gupta et al., 2014 | Accuracy | 0.7 | 5 | -- | sessions 49 |
| Recurrent Neural Network | Hasan et al., 2018 | F1-score | 0.86 | 12 | -- | sessions 129 |
| Markov Chain Model | Hasan et al., 2018 | F1-score | 0.7 | 12 | -- | sessions 129 |
| Hidden Markov Model | Hasan et al., 2018 | F1-score | 0.61 | 12 | -- | sessions 129 |
| Support Vector Machine | Howes et al., 2013 | Accuracy | 0.662 | 20 | -- | sessions 138 |
| J48 | Howes et al., 2013 | Accuracy | 0.51 | 20 | -- | sessions 138 |
| Latent Dirichlet Allocation | Howes et al., 2013 | Kolmogorov-Smirnov | D = 0.300, p = 0.257 | 20 | -- | sessions 138 |
| Random forest Model | Imel et al., 2015 | Accuracy | 0.87 | 4 | -- | 9,300,000 words |
| Boostexter | Lacson and Barzilay, 2005 | Accuracy | 0.73 | 4 | -- | 17,384 words |
| Linear Regression | Malandrakis and Narayanan, 2015 | Pearson Correlation Coefficient | 0.8 | 11 | Strong | sessions 312 |
| Ridge Logistic Regression | Mayfield et al., 2014 | Accuracy | 0.712 | 3 | -- | sessions 415 |
| Ridge Logistic Regression | Mayfield et al., 2014 | Kappa | 0.573 | 3 | Moderate | sessions 415 |
| Random forest Model | Mieskes and Stiegelmayr, 2018 | F1-score | 0.20-0.46 | 7 | -- | sessions 35 |
| Random forest Model | Mieskes and Stiegelmayr, 2018 | Kappa | 0-0.49 | 7 | Poor to Moderate | sessions 35 |
| Support Vector Machine | Perez-Rosas et al., 2017 | F1-score | 0.63-0.84 | 10 | -- | sessions 277 |
| Support Vector Machine | Perez-Rosas et al., 2019 | F1-score | 0.87 | 2 | -- | sessions 151 |
| Automated Co-occurrence Analysis for Semantic Mapping | Salvatore et al., 2012 | Kappa | 0.378 | 14 | Fair | sessions 48 |
| K-Nearest-Neighbours | Sen et al., 2017 | Accuracy | 0.71 | N/A | -- | sessions 122 |
| Recurrent Neural Network | Singla et al., 2018 | F1-score | 0.42-0.60 | 3 | -- | utterances 85,015 |
| Recurrent Neural Networks | Tanana et al., 2016 | F1-score | 0-0.93 | 11 | -- | 1,700,000 words |
| Discrete Sentence Features | Tanana et al., 2016 | F1-score | 0-0.94 | 11 | -- | 1,700,000 words |
| Discrete Sentence Features | Tanana et al., 2016 | Intra-Class Correlation | Excellent = 6 codes Good = 1 codes Fair = 0 codes Poor = 4 codes | 11 | Excellent = 6 codes Good = 1 codes Fair = 0 codes Poo = 4 codes | 1,700,000 words |
| Recurrent Neural Networks | Tanana et al., 2016 | Intra-Class Correlation | Excellent = 6 codes Good: 0 codes Fair = 1 codes Poor = 4 codes | 11 | Excellent = 6 codes Good = 0 codes Fair = 1 codes Poor = 4 codes | 1,700,000 words |
| Discrete Sentence Features | Tanana et al., 2016 | Kappa | Almost Perfect =: 1 codes Substantial = 4 codes Moderate = 1 code Fair = 2 codes Slight = 3 codes | 11 | Almost Perfect = 1 codes Substantial =4 codes Moderate = 1 code Fair = 2 codes Slight = 3 codes | 1,700,000 words |
| Recurrent Neural Networks | Tanana et al., 2016 | Kappa | Almost Perfect = 1 code Substantial = 1 codes Moderate = 3 codes Fair = 3 codes Slight = 3 codes | 11 | Almost Perfect = 1 code Substantial = codes Moderate = 3 codes Fair = 3 codes Slight = 3 codes | 1,700,000 words |
| Markov-Multinomial Model | Wallace et al., 2013 | F1-score | 0.23 | 33 | -- | sessions 476 |
| Joint Additive Sequential Model | Wallace et al., 2013 | F1-score | 0.21 | 33 | -- | sessions 476 |
| Conditional Random Field | Wallace et al., 2014 | Accuracy | 0.64 | 6 | -- | sessions 476 |
| Conditional Random Field | Wallace et al., 2014 | Kappa | 0.47-0.53 | 6 | Moderate | sessions 476 |
| Maximum Likelihood Model | Xiao et al., 2012 | F1-score | 0.56 | 2 | -- | sessions 116 |
| Support Vector Machine | Xiao et al., 2015 | F1-score | 0.89 | 2 | -- | 8,298,507 words |
| Maximum Entropy Markov Model | Xiao, Can, et al., 2016 | Accuracy | 0.85 | 2 | -- | sessions 1,553 |
| Maximum Likelihood Model | Xiao, Can, et al., 2016 | Accuracy | 0.85 | 2 | -- | sessions 1,553 |
| Recurrent Neural Network | Xiao, Huang, et al., 2016 | Accuracy | 0.75 | 8 | -- | 1,414,000 words |
| Maximum Entropy Markov Model | Xiao, Huang, et al., 2016 | Accuracy | 0.72 | 8 | -- | 1,414,000 words |
| Recurrent Neural Networks | Xiao, Huang, et al., 2016b | Kappa | 0.40-0.95 | 8 | Moderate | 1,414,000 words |
| Naïve Bayes | Blanchard et al., 2016b (Semi-automatic...) | F1-score | 0.66 | 2 |  | 1,000 utterances |
| J48 | Blanchard et al., 2016b (Identifying Teacher...) | F1-score | 0.59 | 2 | - | 10,080 utterances |
| J48 | Blanchard et al., 2016b (Identifying Teacher...) | Pearson Correlation r | 0.85 | 2 | Strong correlation | 10,081 utterances |
| Naïve Bayes classifier using the WEKA machine learning toolbox | Donelly et al., 2016 (Multi-sensor...) | F1-score | 0.43-0.51 | 5 | - | 32,134 utterances |
| Naïve Bayes classifier using the WEKA machine learning toolbox | Donnely et al., 2016 (Automatic teacher...) | F1-score | 0.23-0.68 | 5 | - | 2,254 utterances |
| J48 Decision Tree | Donnelly et al., 2017 | F1-score | 0.66 | 2 | - | 10,080 utterances |
| J48 Decision Tree | Samei et al., 2014 | Kappa | 0.24-0.28 | 2 | Fair | 9,579 utterances |
| J48 Decision Tree | Samei et al., 2014 | Accuracy | 0.62-0.64 | 2 | - | 9,580 utterances |
| Decision Tree | Samei et al., 2015 | Accuracy | 0.62-0.68 | 2 | - | 20,737 utterances |
| Random Forest | Wang et al., 2014 | Accuracy | 0.84 | 3 | - | 13 classroom sessions |
| Random Forest | Wang et al., 2014 | Kappa | 0.76 | 3 | Substantial | 14 classroom sessions |
| Bidirectional Long Short Term Memory (Bi-LSTM) | Chen et al., 2019 | F1-score | 0.5876 | 8 | - | sessions 337 |
| Bidirectional Long Short Term Memory (Bi-LSTM) | Chen et al., 2019 | Accuracy | 0.6328 | 8 | - | sessions 337 |
| Bidirectional Long Short Term Memory (Bi-LSTM) | Gibson et al., 2019 | F1-score | 0.619 | 11 | - | 10,844,000 words |
| Bidirectional Long Short Term Memory (Bi-LSTM) | Gibson et al., 2019 | F1-score | 0.784 | 11 | - | 10,844,000 words |
| Logistic Regression | Park et al., 2019 | F1-score | 0.7515 | 27 | - | talk-turns 122,083 |
| Support Vector Machine | Park et al., 2019 | F1-score | 0.745 | 27 | - | talk-turns 122,083 |
| Gated Recurrent Unit | Park et al., 2019 | F1-score | 0.7368 | 27 | - | talk-turns 122,083 |
| Conditional Random Field | Park et al., 2019 | F1-score | 0.6764 | 27 | - | talk-turns 122,083 |
| HMM-LR | Park et al., 2019 | F1-score | 0.7655 | 27 | - | talk-turns 122,083 |
| HMM-SVM | Park et al., 2019 | F1-score | 0.6863 | 27 | - | talk-turns 122,083 |
| HMM-GRU | Park et al., 2019 | F1-score | 0.7706 | 27 | - | talk-turns 122,083 |
| Hier-GRU | Park et al., 2019 | F1-score | 0.7778 | 27 | - | talk-turns 122,083 |
| Support Vector Machine | Flemotomos et al., 2018 | F1-score | 0.86 | 11 | - | sessions 386 |
| Gated Recurrent Unit | Cao et al., 2019 | F1-score | 0.654 | 8 | - | sessions 377 |
| Recurrent Neural Network | Park et al., 2021 | Pearson correlation coefficient | 0.6 | 3 | Moderate | 210,000 utterances |
| Logistic Regression | Park et al., 2021 | Pearson correlation coefficient | 0.55 | 3 | Moderate | 210,000 utterances |
| CNN-BiLSTM | Song et al., 2020 | F1-score | 0.68 | 7 | - | sessions 155 |
| Bi-directional Long Short-Term Memory | Suresh et al., 2019 | F1-score | 0.65 | 8 | - | 60,241 sentences |
| Ridge Regression | Goldberg et al., 2020 | Mean Squared Error, Spearman's rank correlation | MSE = 0.67 Spearman’s p = 0.15, *p* < .001 | 2 | - | sessions 1235 |