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 1.** *Automated Coding Models’ Description*

|  |  |
| --- | --- |
| Model1 | Description |
| Support Vector Machine (SVM) | SVM is a discriminative classifier that is based on the idea of finding a hyperplane that best separates a dataset into two classes. In other words, given the labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In a two-dimensional space, this hyperplane is a line dividing a plane into two parts in which each class lies on either side. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra.  |
| Artificial Neural Networks (ANN) | ANN is an algorithm that mimics how the human brain processes information and consist of input and output layers, as well as (in most cases) hidden layers. This model handles the regression and classification problems without the need to explicitly specify any relationships between the input and output variables. ANN model iterates the feedforward and backpropagation processes to identify the optimal amounts of the weights of the network for a single output so the difference between the predicted and the observed outputs is as small as possible. |
| Markov Chain | Markov chain is a stochastic model describing a sequence of possible events. The process of calculating the probability of each event depends only on the state attained in the previous event and not the sequence of states. |
| Hidden Markov (HMM) | HMM is an evolved version of the Markov chain model and assigns labels to each unit in a sequence that are observable or not observable in the world. However, some events such as part-of-speech tags or acoustic events are not observable in the world, so they are called “hidden”. HMM computes a probability distribution over possible labels and chooses the best label sequence. |
| Maximum Entropy (MaxEnt) | MaxEnt (also known as multinomial logistic regression) is a machine learning framework and belongs to a family of classifiers known as “exponential” or “log-linear” classifiers. This model is used for sequence labelling or classification (i.e., assign labels to each event in some sequences). This model is also capable of assigning a weight to particular events.The most common MaxEnt classifier is maximum entropy Markov model.  |
| Maximum Entropy Markov (MEMM) | MEMM is a graphical model for sequence labelling that combines features of hidden Markov (HMM) and maximum entropy (MaxEnt) models. MEMMs model are applied in Natural language processing, specifically in part-of-speech-tagging and information extraction. |
| Decision Tree and J48 | A decision tree is a hierarchical decision model and consist of “if” and “else” questions asked in each node. Eventually, these questions and the path will lead to a predicted class or a continues real-valued outcome.J48 (C4.5) is an algorithm used to generate a decision tree prediction model. This algorithm was developed after ID3 (which is also a decision tree classifier) by Ross Quinlan. The important features of this model are the prediction of discrete and real-valued outcomes, managing missed-values in the input dataset, the prunning ability to prevent overfitting, and weighting features.  |
| Random Forest | Random Forest model is an ensemble learning model for regression, classification and other tasks. This model applies some decision tree models to predict the outcomes and uses the most common prediction or mean of the predictions as the final outcome prediction for each observation.  |
| Conditional Random Field (CRF) | CRFs are a class of statistical modelling method used to predict sequences rather than discrete or real-valued outcomes. This model is best suited to prediction where contextual information or state of the neighbours affect the current prediction. |
| Labelled Topic | A labelled Topic model is a statistical model that discovers the abstract topics that occur in a series of documents. This model is mainly used in natural language processing and text mining applications. |
| Latent Dirichlet Allocation (LDA) and DiscLDA | LDA is a generative probabilistic model used for topic modelling purposes. This model is capable of discovering the hidden topics in a corpus, classify documents based on those topics and summarise corpora in terms of the topics identified. DiscLDA is a discriminative variation on latent Dirichlet allocation (LDA) model in which a class-dependent linear transformation is introduced on the topic mixture proportions.  |
| Maximum Likelihood | The goal of the maximum likelihood model is to find the optimal way to fit a distribution to the data. Based on the data, the distribution can be normal, exponential, gamma or other distributions. This model estimates the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.  |
| AdaBoost | AdaBoost (adaptive boosting) is a type of ensemble learning method which uses an iterative approach to learn from the mistakes of weak classifiers to build a stronger learning algorithm. The basic classifiers could be any classifier, from decision trees to logistic regression. |
| Automated Co-occurrence Analysis for Semantic Mapping (ACASM) | ACASM constructs a map of the text in terms of thematic nuclei active in it. It works through invariant, ostensible, yet context-sensitive procedures, defined in terms of computational algorithms. |
| Boostexter | BoosTexter is a text-mining tool that uses a machine-learning technique named boosting (using variations of AdaBoost algorithm). This model categorises a text corpus by combining many simple and moderately inaccurate categorization rules into a single, highly accurate categorization rule. |
| Discourse Flow Analysis (DFA) | DFA is a technique developed specifically for the psychotherapy domain and focuses on temporal patterns of meanings rather than on the survey of discrete contents. It also considers the contextual features. |
| Discursis software | Discursis is an automated computer visualisation measurement software that is used to analyse conversational behaviour. Discursis automatically builds an internal language model from a transcript, mines the transcript for its conceptual content, and generates an interactive visual account of the discourse. The resultant visual account of the whole consultation can be analysed for patterns of engagement between interactants. |
| Fidelity Automatic RatEr (FARE) | FARE is a computational system that uses a transcribed text as input, then applies a decision tree algorithm that categorizes linguistic patterns associated with high or low fidelity. |
| K-Nearest Neighbors | K-nearest-neighbours is a simple algorithm that classifies events based on a similarity measure (e.g., distance functions). This model classifies the events based on the class that is most common among K neighbours of that event. K is decided by the researcher.  |
| Linear Regression | Given a set of observations, each observation is associated with some features. Linear regression model is used to predict some real-valued outcome for each observation. The predictive power of this model is boosted when more than one feature is used (in this situation, the model is called multiple linear regression).  |
| Logistic Regression and Lasso Logistic Regression and Ridge Logistic Regression | Given a set of observations, each observation is associated with some features. The Logistic regression model is used to predict some discrete outcomes such as classes. For example, in a binary classification of cancer detection, considering some features, the model outcomes are two classes: “positive” or “negative”. In case the model is used to predict many discrete outcomes, the model is called “multinomial logistic regression”.Least absolute shrinkage and selection operator (LASSO) logistic regression model is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.In Ridge logistic regression model, variables with minor contribution have their coefficients close to zero. However, all the variables are incorporated in the model. This is useful when all variables need to be incorporated in the model according to domain knowledge. |
| Naive Bayes | Naive Bayes model is a classification method based on Bayes theorem. This model assumes that the predictors of a feature are independent from each other. This classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The three variations of this model are Gaussian, multinomial, and Bernoulli. Gaussian assumes that the features follow a normal distribution, Multinomial model is used for classification problems with discrete features (e.g., word counts for text classification). In the Bernoulli model, the features are assumed to be a binary-valued (Bernoulli, Boolean) variable. |
| RapidMiner | RapidMiner is a text mining program. This program is suitable for those who are not interested in programming or simply prefer to use the existing software. Developers of this program tried to integrate various operations in the field of data science which allows researchers to quickly apply it for data mining operations. |

*Note.* 1We described the main methods in this table. ُSome of the included papers used variations of these methods. We reported the specified methods used in each paper in Table 1, Automated Coding Method column.

Descriptions of Coding models’ Accuracy Measures

## **Cohen’s kappa**

Cohen’s kappa coefficient (k) is used for measuring the agreement between human coders and a coding model’s prediction. It is computed as (observed accuracy—expected accuracy)/( 1—expected accuracy). Cohen’s kappa value ranges between 0-1. Kappas < .40 are considered “fair” to “poor,” .41-.60 are “moderate,” .61-.80 are “substantial,” and > .81 are “almost perfect” (Landis & Koch, 1977).

## **Inta-Cater Correlation Coefficient**

The intraclass correlation coefficient (ICC) is a measure of the reliability of measurements or ratings. This measure shows the agreement between a human coder and a model on a session-level prediction. An ICC < .40 is considered a poor level of agreement, ICC between .40-.59 is a fair agreement, ICC between .60-.74 is a good agreement and ICC between .75-1.00 is considered an excellent level of agreement (Cicchetti, 1994).

## **Confusion Matrix**

Confusion Matrix describes the complete performance of the model and presents the exact number of the codes that a model predicted correctly or incorrectly.

|  |  |
| --- | --- |
|  | Actual Class |
| Positive | Negative |
| Predicted Class | Positive | True Positive (TP) | False Positive (FP) |
| Negative | False Negative (FN) | True Negative (TN) |

## **Receiver Operating Characteristic (ROC)**

The receiver operating characteristic (ROC) curve is a plot which shows the performance of a binary classifier as a function of its cut-off threshold. It essentially shows the true positive rate (TPR) against the false positive rate (FPR) for various threshold values.

## **Area Under the Curve (AUC)**

AUC calculates the area under the ROC curve, and therefore it is between 0 and 1. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC can be ranged between 0 to 1. An AUC=0.5 indicates a prediction better than chance level, .50 < AUC < .70 is a poor prediction, .70 =< AUC < .80 is an acceptable level of prediction, .80 =< AUC < .90 is an excellent prediction and AUC >= .90 is an outstanding prediction (Hosmer et al., 2013).

## **Accuracy**

Accuracy is one of the easiest predictive measures and can simply be calculated as the proportion of correctly classified codes (TP + TN) to all the predicted codes (TP + FP + FN + TN). Accuracy is a good measure when the target variable classes in the data are nearly balanced. Accuracy is a good measure when the target variable classes in the data are nearly balanced.

## **Precision**

Precision is a measure that shows what proportion of “positive” predicted codes are actually positive. It is calculated by the number of True Positives divided by the total number of the “Positive” predicted codes (TP+FP).

## **Recall or Sensitivity**

Recall or sensitivity shows what proportion of actual positives are predicted correctly. It is calculated by the number of true positives divided by the truly predicted positives and falsely predicted negatives (TP + FN).

## **Specificity**

Specificity is a measure that shows what proportion of actual negatives, were predicted by the model as negative. It is calculated by this formula: TN/FP + TN. Specificity is the exact opposite of recall.

## **F1-Score**

F1-score is the harmonic mean of precision and recall and is calculated using this formula: F1-score = Precision + Recall/2 \* Precision \* Recall.

## **References**

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment, 6*(4), 284-290.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Landis, J. R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159-174.