A Text analysis labels

The Status category reflects an author's focus on his wealth and occupation, Physical Appearance is associated with an author's focus on physical characteristics (such as length, sexuality, and physical build), and Positive Emotion is related to the author's expressions of positive sentiment. Furthermore, self-references by the author, references to the reader of the dating profile, and references of the author to the reader and them together, are covered by the I, You, and We categories, respectively.

The Status category was formed by joining the LIWC categories Job and Money (1,233 words). The LIWC categories Body and Sexuality together formed Physical Appearance (523 words). The Positive Emotions category only included words from LIWC's Positive Emotions category (1,226 words). The proportion of I-, You-, and We- references were measured using LIWC's I (11 words), You (20 words), and We (7 words) categories.

B Preprocessing and model selection

Stop word removal, named entity removal, removal of frequent words (appearing in more than 25% of texts, with the exception of I-, you-, and we-references) and infrequent words (words appearing less than 5 times total) was applied to the dating profiles.

For the label assignment task, eight regression algorithms were compared (SVR, SGDRegressor, BayesianRidge, LassoLars, PassiveAggressiveRegressor, TheilSenRegressor, LinearRegression, and an 8-layer LSTM described below), TheilSenRegressor was chosen because it returned the highest scores (Dang et al., 2008). Word uni-grams, bi-grams, and tri-grams were used as features, and the labels obtained by human evaluation were used for training.

For the relationship goal identification task, seven algorithms were compared (Linear SVM, Naive Bayes, C4.5, AdaBoost, Random Forrest Classifier, XGBoost, LSTM). The LSTM was chosen for the LIWC, word-based, and Meta classification models as it led to the highest accuracy scores (Hochreiter and Schmidhuber, 1997). The architecture of this model is based on previous (binary) writers' intention classification for Quora data (Bai, 2017) and implemented using Keras (Chollet et al., 2015). The preprocessed word features, represented using one-hot vector encoding, served as input for the word-based model. This model consisted of 8 layers. (1) An embedding layer, (2) a dropout layer, (3) a batch normalization layer, (4) a dense layer with relu activation, (5) a second dropout layer, (6) a second batch normalization layer, (7) an LSTM layer, (8) a second dense layer of 1 dimension with sigmoid activation. Furthermore, the model was trained using binary cross-entropy loss and optimized on accuracy using Adam optimizer. The models trained for a maximum of 20 epochs and early stopping was applied with patience 5 based on accuracy. Bayesian optimization was used for hyperparameter optimization of the model (Snoek et al., 2012). The parameters that were tuned were LSTM units, dimensions of the first dense layer, dropout value of the dropout layers, dropout value of the LSTM layer, output dimension of the embedding layer, and batch size (see Table 6 for the applied hyperparameters). The model was similar for LIWC and the meta-classifier, with exception of the embedding layer, and thereby the output dimension of the embedding layer, which were excluded.

Method	LSTM	dense	lstm_drop	drop	output_dim	batch
Word	216	116	0.19	0.16	72	104
LIWC	175	149	0.16	0.16	-	58
Meta	254	100	0.29	0.40	-	22

Table 6: Hyperparameters used for the classifiers in the relationship goal identification task

LDA models were trained using Mallet (McCallum, 2002). Similar to Thompson and Mimno (2018), hyperparameter optimization occurred every 20 intervals after the first 50. Topics were set to be six, the same number as the selected LIWC labels.