Predicting Moral Judgments of Interpersonal Conflicts

EDLD 654: Machine Learning

Fall 2023

Sabreen NoorAli

https://github.com/SabreenNoorAli/final_project_EDLD654.git

Predicting Moral Judgments of Interpersonal Conflicts

Individuals face a variety of conflicts with others (i.e., interpersonal conflicts) every day for reasons ranging from seemingly unimportant issues such as friends deciding what to do together in their free time to more consequential issues such as suspicions of disloyalty. The examples of interpersonal conflicts are virtually limitless. Upon experiencing interpersonal conflicts, people often evaluate their own actions as well as the actions of the person with whom they had the conflict. These evaluations are moral judgments of interpersonal conflicts. The current project aims to predict these moral judgments.

Most studies on moral judgments heavily rely on presenting research participants with hypothetical scenarios in which participants either judge the action/actor in the scenario or are asked what they would do if they were the actor (e.g., Pizarro et al., 2013). However, hypothetical scenarios do not inform us about how judgments are made in real-life conflicts. Therefore, this project focuses on the moral judgments of real-life conflicts. More importantly, the way that an interpersonal conflict is described by the person who experienced it, as opposed to reading a standard scenario, may contain important cues regarding how the parties involved in the conflict are judged. In fact, research on morality shows that everyday language contains important cues depicting one's moral values (Hofmann et al., 2014). Therefore, in this project, I attempt to predict people's moral judgments of themselves and other people using the linguistic cues in text descriptions of interpersonal conflicts.

A predictive model for moral judgments of the self and the other person would be helpful in two ways. First, it would allow people to gain insight into themselves. One could see how others would judge themselves and the other person if others were in their place. Second, the ability to predict one's moral judgments via text would allow researchers and practitioners to see how individuals generally judge themselves and others when they experience a conflict. For these reasons, it could prove helpful for people to understand themselves better and for others such as researchers and recruiters to understand these people better.

Methods

Data

Data for this project come from a series of studies conducted between 2019 and 2021 at University of Oregon. All studies investigate the relationship between perspective taking and moral judgments of interpersonal conflicts. All studies instructed participants to write about at least memorable conflict that they had with another person.

In Study 1, participants (N = 1334) were randomly assigned to four conditions. Then, participants wrote about one memorable conflict that they had with another person in only one of the conditions using at least 10 sentences. After writing about the conflict, participants responded to two items that measured the extent to which they thought they and the other person in the conflict were right (i.e., judgment of self and the other person) on a 7-point scale (1 = Completely Wrong; 7 = Completely Right). Also, participants indicated the extent to which they thought they included the thoughts and feelings of the other person in the conflict (i.e., perceived perspective taking) on a 7-point scale (1 = Not at all; 7 = Completely). Participants were also asked for their permission to use the descriptions of conflicts to be used by the researchers in future studies. Next, demographic information (i.e., age, gender, and race) was collected.

Study 2 (N = 496) had a within-subject design in which participants wrote about two conflicts that they had with another person. Participants first wrote about a conflict in the control condition and then another one in one of the conditions ($N_{observation}$ = 992). After writing about each conflict, participants responded to the same items from Study 1. However, due to a programming mistake, data for race was not collected in this study. Finally, Study 3 ($N = 410$) had the identical design as Study 1, but Study 3 had only two conditions. Since there was no difference between the conditions, this variable was removed from the data.

The data from all three studies were merged. Thus, the total number of observations was 2746. The combined dataset had two outcomes (i.e., judgment of self and the other person) and four features (i.e., perceived perspective taking, age, gender, and race). However, race was removed from the data due to missing data (39.91%). Gender was also removed from the data since most of the sample was female (61.25%).

In addition to the core features, we extracted additional features from the conflict descriptions using several R packages. Basic text statistics (e.g., number of sentences), text entropy, lexical diversity measures (e.g., Herdan's *C*), and readability measures (e.g., average familiar words) were extracted using the *quanteda* package (Benoit et al., 2018). These analyses provided 73 features. Five word-length statistics (e.g., mean word length per conflict description) were also extracted. Using the *udpipe* package (Wijffels, 2021), part of speech tags (e.g., adjectives), morphological features (e.g., accusative), and syntactic relations (e.g., relative causal modifier) were extracted. These analyses provided 147 features. Roberta-base (12th layer) was used to extract 768 features using the *reticulate* (Ushey et al., 2021) and *text* (Kjell et al., in progress) packages. We also used the Linguistic Inquiry and Word Count software to extract 89 features (e.g., personal concerns; ref). Finally, using the *quanteda.dictionaries* package (Benoit et al., 2018), we extracted 10 features that represent both positive and negative dimensions moral foundations (e.g., care virtuous; Graham et al., 2013). Thus, along with the core features, the final dataset included 1095 features in total. See online supplementary materials (i.e., feature_generation.html) for the list of the features as well as descriptive statistics for each feature.

Before fitting the models, a blueprint was created such that all numeric predictors that have zero or near zero variance were removed, the missing values in numeric predictors were mean-imputed, all the numeric predictors were standardized, the predictors that are correlated at greater than .8 were removed (one of them was removed), and the other outcome was also removed. Then, this blueprint was passed on to the model. Finally, 80% of the data were used as the training data, and the rest was the test data. In the training data, I used a 10-fold cross-validation procedure.

Models

The models were fit using the packages *recipes* (Kuhn & Wickham, 2021), *caret* (Kuhn, 2021), *glmnet* (Friedman et al., 2010), and *gbm* (Greenwell et al., 2020). Due to the high number of predictors in the dataset, I first started with regularized linear regression techniques. In addition, given their relatively comparable performance with regularized regression, I also fit a gradient boosting trees model. Thus, linear regression with ridge penalty, linear regression with lasso penalty, and gradient boosting trees were fit for each outcome (i.e., judgment of self and other person).

The ridge regression is a penalized multiple-regression method that adds a penalty term (i.e., lambda) to the loss function of linear regression. The loss function in ordinary least-squares regression is sums of squared errors (SSR). In ridge regression, the squared of each coefficient multiplied by the lambda is added to the SSR. The lambda is a parameter that penalizes the coefficients when they get larger. As the value of lambda increases, so does the degree of penalty. Therefore, the ridge regression is especially useful when there is a high number of predictors and multicollinearity. Given that there are 1095 predictors in the current project, this technique is particularly helpful. The lambda can take any positive value. Therefore, the lambda is a parameter that needs to be tuned.

In the current project, I started by fitting models for every lambda value between .01 and 3 with increments of .01 using RMSE as the evaluation metric for the judgments of the self. This analysis showed that a lambda value of 3 led to the smallest RMSE value. Thus, I tried even higher values to investigate whether a bigger lambda value would produce a smaller RMSE value than the previous model. I tried lambda values between 3 and 20 with increments of 1. This analysis showed that a lambda value of 7 minimizes the RMSE value. To fine-tune the lambda, I did a final search of lambda values between 6 and 8 with increments of .01. The results showed that a lambda value of 7.08 produced the smallest RMSE. This was the final model that I used to apply to the test data. For the judgments of the other person, the lambda values that I tried were the same as the judgment of the self in the first two attempts to find the optimal lambda value. In my final attempt, I tried lambda values between 3.5 and 4.5 with increments of .01. The best tune was 3.88. Similarly, this model was also applied to the test data.

The lasso regression is also a penalized regression method. However, the absolute value of each coefficient is multiplied by the penalty term (i.e., lambda) as opposed to the squared values as in ridge regression. For both outcomes, I started by trying lambda values between .01 and 3 with increments of .01. The best lambda values that minimized the RMSE were .06 and .04 for the judgment of the self and the other person, respectively.

As a third model, I fit a gradient boosting trees model. The gradient boosting trees are fundamentally decision tree models in which predictors are split into two at certain values to predict the outcome. For instance, values below the mean of a feature predict certain values of outcome while values above the mean predict the rest of the values in the outcome. This process is repeated many times for each predictor and constitutes a tree-like structure to predict the outcome. In certain decision tree models like bagged trees and random forest models, random samples of rows and features are used to create the trees. However, in these models, each iteration is independent from one another. In gradient boosting trees, the trees are developed sequentially. That is, each tree model is built upon the former models' prediction errors. This process usually results in better performance compared to the bagged trees and random forests.

In building gradient boosting models, certain parameters need to be tuned just like in regularized regression and other tree models. However, due to relatively high number of parameters that need tuning, I tried to tune each parameter in a sequential manner for both outcomes. First, I tried to find the best value for the number of trees to have in the model by trying values between 1 and 1000. The optimal number of trees were 21 and 72 for the judgment of the self and the other person, respectively. Next, I tried to find the best value for the maximum depth of each tree developed at each iteration (i.e., interaction depth) and the minimum number of observations in each terminal note of tree models at each iteration. For this reason, I did a grid search for every combination of interaction depth between 1 and 15 and the minimum number of observations in each node between 5 and 50 with increments of 5. The number of trees were fixed at the values detected in the first step. The results showed that the interaction depth of 7 and 13 and the minimum observation numbers of 40 and 50 were the optimal values for the judgment of the self and the other person, respectively. Up to this step, I always fixed the learning rate at 1. The learning rate refers to the ratio of predicted values used for the new predictions. Thus, in the former steps, I used 100% of the predicted values (by multiplying them by 1). However, the lower learning rates generally leads to better prediction in a way that is analogous to the speed-accuracy tradeoff. Therefore, in the final step, I lowered the learning rate to 10% (by multiplying the values by .1) and increased the number of trees to 5000 to get a better model. The interaction depth and the minimum number of observations were fixed at the values found in the second step. The results indicated that 309 and 528 trees were ideal for the judgment of the self and the other person, respectively. Therefore, in the final model for the judgment of the self, the number of trees was 309, the interaction depth was 13, the learning rate was .1, and the minimum number of observations in each node was 40. As for the final model for the judgment of the other person, the number of trees was 528, the interaction depth was 7, the learning rate was .1, and the minimum number of observations in each node was 50.

Finally, I evaluated these models using three criteria. I used mean absolute error (MAE) and root mean squared error (RMSE) to investigate the accuracy of the predictors. I tried to minimize these values. In addition, I computed the R^2 value to evaluate the extent to which the models explain the variance in the outcome, which I tried to maximize.

Results

Table 1 summarizes the performance of the models for each outcome. For the judgment of the self, the gradient boosting trees model provided the relatively best model such that while its MAE and RMSE values were the smallest, its R^2 values was the highest compared to the linear regression with ridge and lasso penalty. For this outcome, ridge and lasso regression resulted in close performance. However, the ridge regression had a particularly low R^2 value. As for the judgment of the other person, the ridge regression seems to be the slightly better model compared to the other two models. Even though the ridge regression had the highest MAE and RMSE, this difference was very minor, but it explained more than 1% variance compared to the other models. For the other models, it is difficult to say which one is better or worse. The lasso regression had a lower MAE and R^2 but a higher RMSE compared to the gradient boosting trees.

The results for both outcomes indicate relatively smaller R^2 values and relatively greater values of MAE and RMSE than expected. That is, the model for the judgment of the self explained only 7.1% of the variance in the outcome. This value was 5.6% for the judgment of the other person. This was surprising in a sense. That is, my initial expectation was to achieve a better model with lower error rates and higher R^2 values. The sample size was not really small and there was a high number of features. However, it seems that these features do not do well predicting moral judgments of interpersonal conflicts.

Table 1. Performance evaluation metrics across models and outcomes

	MAE	RMSE	R^2
Judgment of the self			
Ridge regression	1.2916	1.6011	0.00002
Lasso regression	1.3011	1.6009	0.0011
Gradient Boosting Trees	1.2018	1.5061	0.0712
Judgment of the other person			
Ridge regression	2.6682	3.0204	0.0565
Lasso regression	2.6475	3.0037	0.0390
Gradient Boosting Trees	2.6555	2.9964	0.0427

MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Figure 1 demonstrates the observed values in the outcomes and the predicted values by the best models. The depiction of observed and predicted values seems to successfully visualize the gap between the two values. While the observed values seem to have a wider distribution, the predicted values reflect a more continuous distribution around the mean score for each outcome. However, this situation seems to be the result of the way that the outcomes are measured (i.e., Likert scales). This could indicate that if the outcomes were measured on a truly continuous scale, the predicted values would overlap more with the observed values, which would also lead to better model fit statistics.

Figure 1

Observed and predicted values of the outcomes

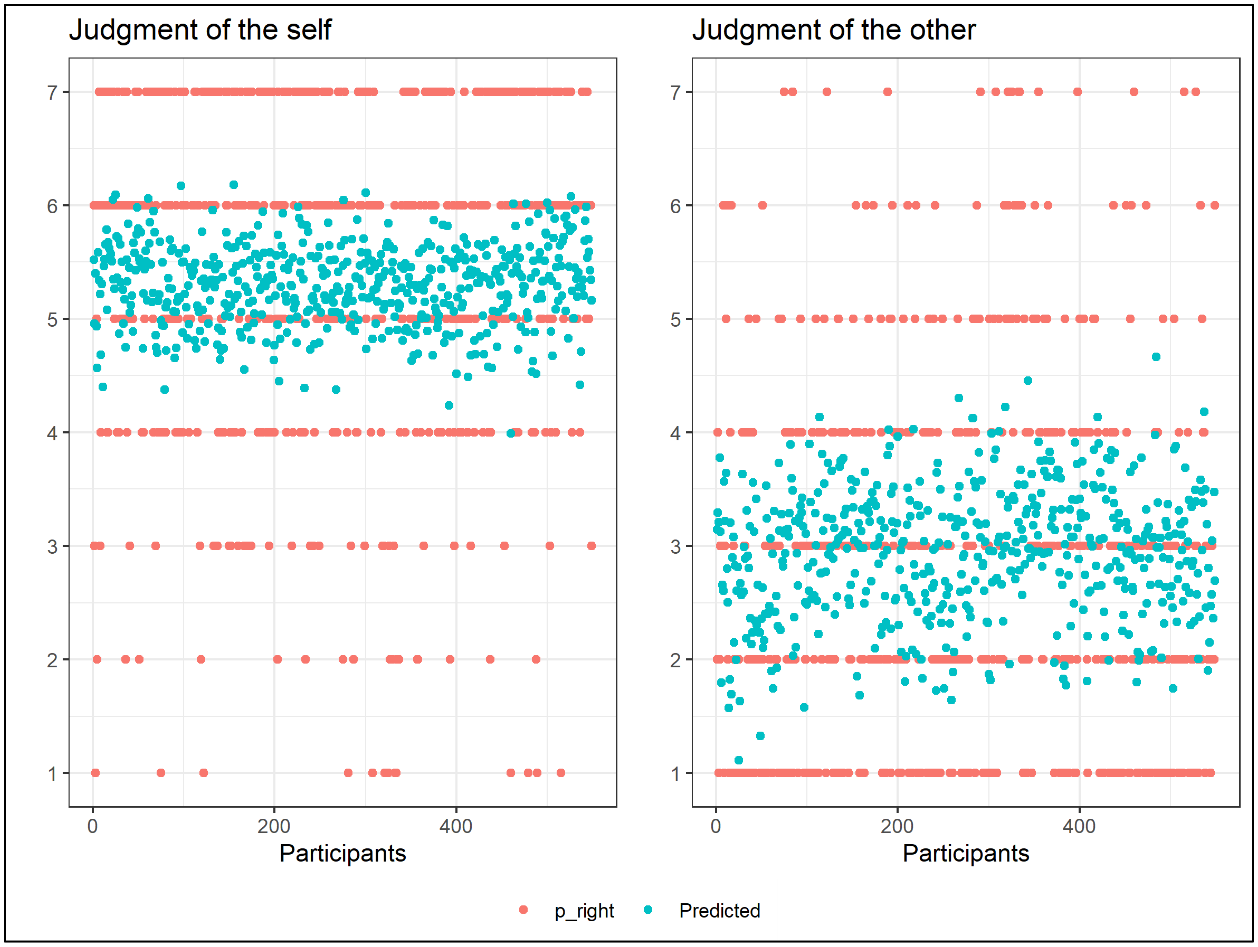
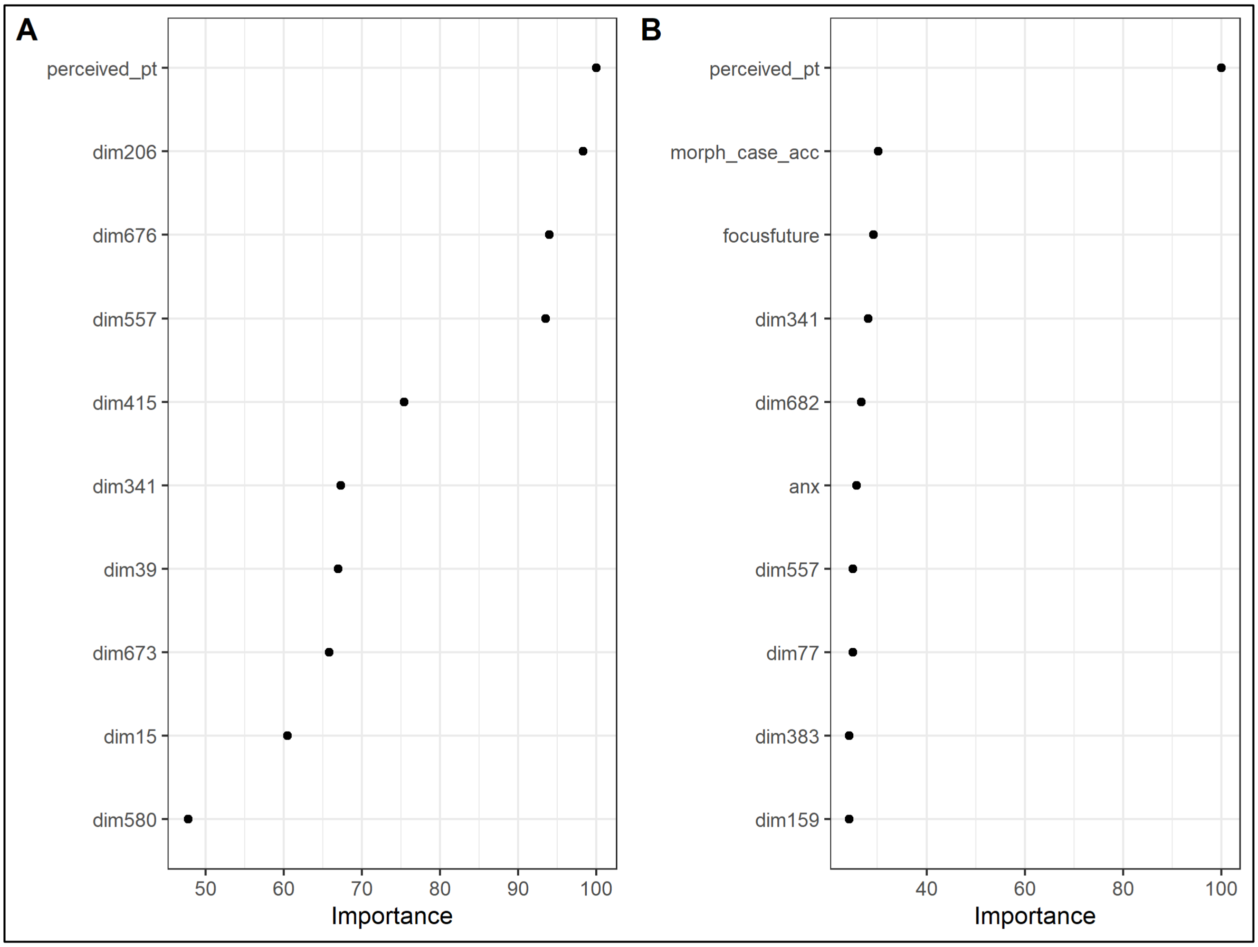


Figure 2 visualizes the most important 10 predictors in each outcome. The perceived perspective taking feature was found to be the important predictor for both outcomes. For the judgment of the self, the rest of the important features were the ones extracted from the text using natural language processing (i.e., Roberta-base). As for the judgment of the other person, the rest of the important features were again mostly the ones extracted from the text using natural language processing, but there were also some morphological features as well as one feature extracted using LIWC.

Figure 2

The most important predictors in ridge regression models for both outcomes



Note. Figure A represents the judgment of the self. Figure B represents the judgment of the other person. Best models were used for the visualizations of feature importance.

Discussion

The current project aimed to develop a predictive model for the moral judgments of self and other person in interpersonal conflicts. I fit three models for each outcome and evaluated the accuracy of the predictors and the amount of variance explained in the outcome by the predictors for each model. The results showed that the gradient boosting trees was the best for the judgment of the self, and the linear regression with ridge penalty was the best model for the judgment of the other person. However, the performance of each model was quite close to one another for both outcomes. The best models for both outcomes only slightly outperformed the other models.

Regarding the models for the judgment of the self, I should also note that the learning rate as well as the bag fraction were not tuned in the gradient boosting trees models due to their computational cost. Therefore, it is unclear whether the gradient boosting trees would perform better had these hyperparameters been tuned. Nevertheless, with limited resources, the regularized regression techniques seem to be desirable since they are significantly less computationally expensive than the gradient boosting trees. Even without tuning these hyperparameters, the gradient boosting trees outperformed the penalized regression methods for the judgment of the self. This was a surprising result since the regression methods usually outperform tree models for continuous outcomes.

The number of predictors were quite high considering the sample size. For this reason, I also identified the most important predictors in the final models. For the judgment of self, the most important predictors included perceived perspective taking and 9 dimensions extracted via natural language processing. It was not surprising to see the perceived perspective taking as the most important variable since the data for this project came from a study about the effect of perspective taking on moral judgments. However, I did not expect the features extracted by the natural language processing to be more important than the other features. I was expecting to see at least a few features extracted using the moral foundations dictionary since the outcomes are directly related to the moral values. As for the judgment of the other person, the most important predictor was also perceived perspective taking. This is consistent with the findings for the other outcome. Similar to the findings of the judgment of the self, most of the important predictors were some of the features extracted by the natural language processing. Interestingly, dimension 557 was an important predictor for both outcomes. There were also a couple of morphological features and one feature from the LIWC library as important predictors. I would expect to see more features extracted using other methods as well.

One of the most striking results was the amount of variance explained. The R^2 values were consistently very low across the models and outcomes. The models were able to explain 4-7% of the variance in the outcomes. Coupled with the relatively low accuracy of the predictors, the poor performance reduces the models' theoretical and practical significance. Thus, it is unclear how models that cannot explain most of the variance in the outcomes can inform the theory or be helpful in a practical way. The reasoning behind this project was to create a predictive model that can help one see how people generally think about themselves and others when they experience an interpersonal conflict. Given the findings, it would be unhelpful to recommend any of the models as a tool for this purpose.

The low performance of the models may be because the moral judgments of self and other person are difficult to predict based on textual descriptions of interpersonal conflicts. There may be a variety of psychological and situational factors that contribute to one's evaluation of themselves and others with whom they had a conflict. In this case, more features that get at such factors would increase the performance of the models. However, these features are usually not readily available in the text data. In this context, future research should focus on what can be predicted using text data. For instance, moral judgments of the actors involved in an interpersonal conflict by third-party observers would likely lead to better performance since the only data available to third-party observers would be the description of the conflict.

Finally, judging by the visualization of the observed and predicted values, it is likely that the observed values may not be at interval or ratio scale. A model in which the outcome is considered ordinal such an ordinal regression model may yield better performance compared to linear regression models tested in this project where the outcome is assumed to be measured on interval or ratio scales.

References

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software, 3*(30), 774. doi: 10.21105/joss.00774.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1-22. URL: https://www.jstatsoft.org/v33/i01/.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Greenwell, B., Boehmke, B., & Cunningham, J., & GBM Developers (2020). gbm: Generalized Boosted Regression Models. *R package version 2.1.8*. https://CRAN.R-project.org/package=gbm

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340-1343.

Kuhn, M. (2021). caret: Classification and Regression Training. *R package version 6.0-90*. https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2018). recipes: Preprocessing tools to create design matrices. *R package version 0.1.17*. https://CRAN.R-project.org/package=recipes

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological science, 14*(3), 267-272.

Ushey, K., Allaire, JJ., & Tang, Y. (2021). reticulate: Interface to 'Python'. *R package version 1.22*. https://CRAN.R-project.org/package=reticulate

Wijffels, J. (2021). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. *R package version 0.8.6*. https://CRAN.R-project.org/package=udpipe