Mapping Individual Semantic Spaces using the Demonstrative Choice Task

Anders Havbro Hjulmand (201910585)

Supervisor: Mikkel Wallentin

Bachelor Thesis in Cognitive Science

School of Communication and Culture, University of Aarhus

January 4, 2023

Abstract

Demonstrative choice (e.g. *this* or *that*) is influenced by a range of factors that emerges from the relationship between the speaker and the referent. These include physical distance, ownership, familiarity, valence, and the self. While previous studies have focused on factors that are common among participants, individual differences in demonstrative use remain yet to be systematically investigated. This study used the Demonstrative Choice Task (DCT) to map the position of 480 words in the semantic space of 3,014 participants. After pairing words with demonstratives, participants completed short questionnaires for depression, anxiety, and Big Five personality traits. It was found that the DCT successfully mapped individual differences in semantic landscapes. Depressed and anxious individuals used more proximal demonstratives (*this*) for words that evoked fear, sadness, and disgust, and they used more distal demonstratives (*that*) for positive words. Furthermore, participants' response patterns in the DCT were found to predict depression and anxiety. Many statistical models generalized to novel data, indicating that similar results are likely to be replicated in subsequent studies. These findings suggest that demonstrative choice is influenced by the way experiences, memories, personality traits, mental health, etc. shape the individual semantic landscape. Demonstratives can therefore reveal core aspects of individual semantic knowledge, underpinning their role as a powerful interface between language, cognition, and mental illnesses. Carefully selecting the set of words in future studies may increase the predictive power of the DCT.

Keywords: DCT, demonstratives, machine learning, depression, anxiety, Big Five

Character count: 47,844

Code & Data availability: https://github.com/ah140797/Bachelor

Contents

1 Introduction

1.1 Foundations of demonstratives

Demonstratives are deictic expressions that point to a specific referent by providing instructions on how to locate the referent among various referents (Diessel, 2013). In everyday language use, when a speaker says: "could you hand me *that*" or "could you take a look at *this*", the addressee typically identifies the intended referent based on the perceptual context and multimodal cues such as pointing or gaze cuing (Cooperrider, 2016). Demonstratives can therefore establish joint attention on virtually any object, thereby serving one of the most fundamental roles in communication and language (Diessel, 2006, 2014).

The use of demonstratives as efficient communication tools relies on the establishment of a particular point of reference, the deictic center, also called the origo. The origo is the center of a coordinate system from where the world is evaluated. The interpretation of *this* and *that* is determined by the referent's position in relation to the origo. The origo is most often grounded egocentrically in the speaker's body, meaning that demonstratives are closely tied to the speaker's gestures and physical self (Diessel, 2013, 2014).

Furthermore, demonstratives constitute a unique class of linguistic expressions. First, they are one of the few elements of language that are truly universal (Diessel, 2013; Diessel & Coventry, 2020). Second, they are among the most frequent words in the lexicon of children (Clark & Sengul, 1978; Diessel, 2006) and adults (Geoffryy, 2016). Lastly, there is evidence that demonstratives are primordial elements in language evolution (Diessel & Coventry, 2020).

1.2 The choice of demonstrative: *this* or *that*?

It has been widely debated which factors influence the choice of demonstrative. What determines whether a speaker will choose *this* (proximal demonstrative) or *that* (distal demonstrative) when pointing to a referent?
Several studies found that objects in peripersonal space (within reach) evoke more use of *this* and as objects move across the boundary to extrapersonal space (outside reach) there is an increase in the use of *that* (Caldano & Coventry, 2019; Gudde et al., 2016). This effect was also found in Spanish (Coventry, Guijarro-Fuentes, et al., 2008). Moreover, a series of experiments found that the boundary between peripersonal- and extrapersonal space is flexible and graded (Berti & Frassinetti, 2000; Coventry, Valdés, et al., 2008; Farnè et al., 2005; Longo & Lourenco, 2006). These findings suggest that demonstrative choice is affected by the physical distance between the origo and the referent.
Some accounts have attempted to explain demonstrative use by the distance parameter alone (Diessel 2005; Maes, 2007). However, it turns out that accounting for demonstrative use by a single catch-all parameter, such as physical distance, is reductive and incomplete. The profile of demonstrative usage is more nuanced and influenced by a variety of other factors (Peeters et al., 2021).

A study by Rocca and colleagues found that participants transposed the origo onto a partner during a spatial collaboration task, suggesting that demonstrative use is also modulated by the social context (Rocca, Wallentin, et al., 2019; Diessel & Coventry, 2020).
Moreover, a study by Coventry et al. (2014) reinforced the mapping of distance and demonstrative choice, but also provided evidence that the speaker's feeling of ownership and familiarity towards referents influence demonstrative choice.
Furthermore, Rocca and colleagues found that objects which offer more affordances for manipulation (small, harmless, and inanimate) elicited more proximal demonstratives than their

non-manipulative counterparts (large, harmful, and animate), thereby suggesting a link between demonstrative use and object semantics (Rocca, Tylén, et al., 2019).
These findings were extended in another study by Rocca & Wallentin (2020), who found that demonstrative choice was linked to a set of semantic features such as valence, loudness, motion, manipulability, arousal, and the self. Similar results were for the Spanish demonstratives, indicating that the influence of semantic factors on demonstrative choice could be consistent across languages (Todisco et al., 2021).

These studies demonstrate that demonstrative use is shaped by factors that encompass the physical space (distance, visibility) and the semantic space of the speaker (ownership, familiarity, manipulation, valence, the self, etc.). The semantic space revolves around the emotions, memories, motivations, needs etc. that the speaker experiences in relation to a referent.
Evidence suggests that semantic factors might have a greater influence on demonstrative choice than physical factors in face-to-face conversation (exophoric use) (Peeters et al., 2021; Rocca & Wallentin, 2020). These observations lend support to the idea that the origo is centered, not only in the speaker's physical self but also in the speaker's psychological, semantic, and imaginative self, in what Bühler (1934) refers to as "deixis am Phantasma" (Diessel & Coventry, 2020).

Interpreting the origo as anchored in a multidimensional self creates an interesting link between demonstrative use and the semantic space of the speaker. Rocca & Wallentin (2020) hypothesized that the proximal/distal contrast in demonstrative use relates not only to the physical distance between the speaker and the referent, but also to the distance in a multidimensional semantic space. The choice of demonstrative may thus reveal the referent's position in the semantic space of the speaker. Referents that are in close semantic proximity to the speaker could elicit more proximal demonstratives and referents further away could elicit more distal demonstratives.

While many studies have examined the link between demonstrative use and semantics on a general level, individual differences in demonstrative choice remain yet to be systematically investigated. However, it is likely that the same referent can elicit different responses in different individuals. For example, an experienced football player might use proximal demonstratives for familiar words such as *corner kick* or *red card*, whereas a person who is ignorant about football might use distal demonstratives for the same words. Demonstrative choice may thus be affected by the way experiences, memories, personality traits, mental health, etc. shape the individual semantic landscape (Peeters et al., 2021; Rocca Wallentin, 2020).

1.3 The present study

This study investigated whether demonstratives could map individual differences in personality. In the experiment, participants were presented with words and instructed to choose between the proximal or distal demonstrative without further context (endophoric use), in what is known as the Demonstrative Choice Task (DCT). The DCT has been found to be a reliable method for examining the position of words in the semantic space of participants, making it a suitable experimental paradigm for this study (Rocca & Wallentin, 2020).
After pairing words with demonstratives, participants were given screening questionaries for depression, anxiety, and Big Five personality traits, where the trait of emotional stability was of particular interest. Depression is characterized by symptoms such as fatigue, loss of interest in normal activities, a feeling of sadness/hopelessness, and sleep disturbances. Anxiety is characterized by a persistent feeling of worry and fear, increased heart rate, and trouble relaxing (Tiller, 2013). Emotional stability is a fundamental personality trait, and a lack of emotional stability is characterized by a frequent feeling of anxiety and a lack of self-confidence (Ellis et al., 2018, p. 76). Given these characteristics, it was expected that the semantic landscapes of

participants with depression, anxiety, and low emotional stability were different from a healthy landscape.

First, it was hypothesized that demonstratives could map the differences between the semantic landscapes of healthy participants and participants with depression, anxiety, and low emotional stability according to the proximal/distal contrast. The semantic landscape of interest was defined by a set of eleven core emotional dimensions: *Valence*, *Arousal*, *Dominance*, *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, and *Trust*. These dimensions were chosen as they arguably constitute core emotions (Plutchik, 1994; Russell, 2003), and because depression, anxiety, and emotional stability are characterized by changes in the emotional aspect of the self (Tiller, 2013). It was therefore expected that differences in semantic landscapes could be particularly salient in these specific emotional dimensions, although they are not an exhaustive characterization of the semantic landscapes.

Second, it was hypothesized that participants' response patterns in the DCT could predict depression, anxiety, and emotional stability. Moreover, it was expected that the difference in demonstrative use would be salient only in a subset of the words. For example, words such as *lonely* and *danger* might elicit more proximal demonstratives in individuals with depression, anxiety, and low emotional stability, whereas words such as *bread* and *farmer* might not elicit any difference in demonstrative use. The effectiveness of the DCT in classifying depression, anxiety, and emotional stability was therefore expected to be dependent on the words in the experiment.

The analysis in this study emphasizes the predictive framework of machine learning. The predictive framework is used because results from an explanatory framework are often not replicated in subsequent experiments, meaning that models only capture phenomena that exist in the limited amount of data in the experiment, what is referred to as overfitting. In contrast, models from a predictive framework are evaluated on their ability to predict on novel datasets. Results from a predictive framework are therefore more likely to capture the true phenomena of the data, thereby providing robust models that generalize to new contexts (Yarkoni & Westfall, 2017).

2 Methods

2.1 Participants

The study included a total of 3,014 participants recruited through the Prolific website[1] where 2,986 were native English speakers, and 28 were non-native English speakers; 1,893 were female, 1,105 were male, and 16 were another gender; 1,230 were 18-30 years, 935 were 30-40 years, 455 were 40-50 years, 315 were 50-60 years, and 79 were 60+ years. Participants received 0.86 GBP for participation and were anonymized.

2.2 Procedure

The experiment was conducted using the Qualtrics website[2] where a total of 480 words were divided into 8 subsets of 60 words each. The participants were randomly allocated to one such subset of 60 words, yielding a total of 8 groups with the following number of participants in each group: n(1)=362, n(2)=378, n(3)=411, n(4)=369, n(5)=358, n(6)=395, n(7)=366, n(8)=375.

[1] https://www.prolific.com
[2] https://www.qualtrics.com

Each participant was instructed to choose between the demonstrative *this* or *that* for each word, without further context. The lack of context was meant to rule out possible confounds (social, spatial etc.) so that the choice of demonstrative would be determined by the position of the word in the semantic space of the participants.
After pairing words with demonstratives, the participants completed three questionaries: Patient Health Questionnaire (PHQ-9) which is a nine-item screening tool for depression (Kroenke et al., 2001); General Anxiety Disorder-7 Assessment (GAD-7), which is a seven-item screening tool for anxiety disorders (Spitzer et al., 2006); and Ten-Item Personality Inventory (TIPI) which is a measure of the Big Five personality traits containing two items per dimension, where the trait of emotional stability was of particular interest (Gosling et al., 2003). All the questionnaires were extremely brief to ensure high quality of participants' responses.

There was a total of 180,840 trials in the study. Each participant took an average of 8 minutes and 43 seconds to complete the experiment. The study was approved by the Institutional Review Board at Aarhus University.

2.3 Factor analysis

In the present study, there were moderate to high correlations between measures of depression, anxiety, and emotional stability (see top section of figure 1a). This pattern is also found in clinical settings, where about 85% of patients with depression have significant anxiety, and 90% of patients with anxiety disorders have depression (Tiller, 2013). Despite their comorbidity, Spitzer et al. (2006) argue that there is value in assessing depression and anxiety as two separate dimensions.

A factor analysis was conducted on the three questionaries, particularly on all nine items from PHQ9 measuring depression; all the seven items from GAD7 measuring anxiety; and the two items from TIPI measuring emotional stability, yielding a total of 18 items. The two items from TIPI were reverse scored to be consistent with scores from PHQ9 and GAD7, such that high scores corresponded to less emotional stability.
The aim of the factor analysis was twofold: First, to reduce the dimensionality of the three measures, and second to produce new measures that were less correlated than the original measures. The factor analysis was conducted in `R v.4.2.1` (R Core Team, 2021) using the package `psych v.2.2.5` (Revelle, W, 2022). The following paragraphs outline the procedure of the factor analysis.

First, two tests were conducted to determine the suitability of a factor analysis. Bartlett's Test of Sphericity was highly significant ($\chi^2(153) = 36161$, $p < 0.001$), demonstrating that the correlation matrix of the 18 items was different from the identity matrix (18×18 matrix with 1's on the diagonal and 0's on the off-diagonal). The Kaiser-Meyer-Olkin Test (KMO) indicated that the overall strength of the relationship among the items was very high (KMO = 0.96). Moreover, the pairwise Pearson correlations of the 18 items were in the moderate to high range (see figure 1b). These observations provide strong evidence that the 18 items were indeed suitable for a factor analysis (Williams et al., 2010).

To determine the number of factors, Horns parallel method was used (Horn, 1965). This method randomly shuffled the correlation matrix from the 18 items and conducted a factor analysis. Then, factors whose real eigenvalues were higher than the corresponding eigenvalue from the shuffled data were extracted. The estimated number of factors was three. However, a factor analysis with three factors failed to produce interpretable results and also returned many items with high cross-loadings. Thus, two factors were chosen for the factor analysis.

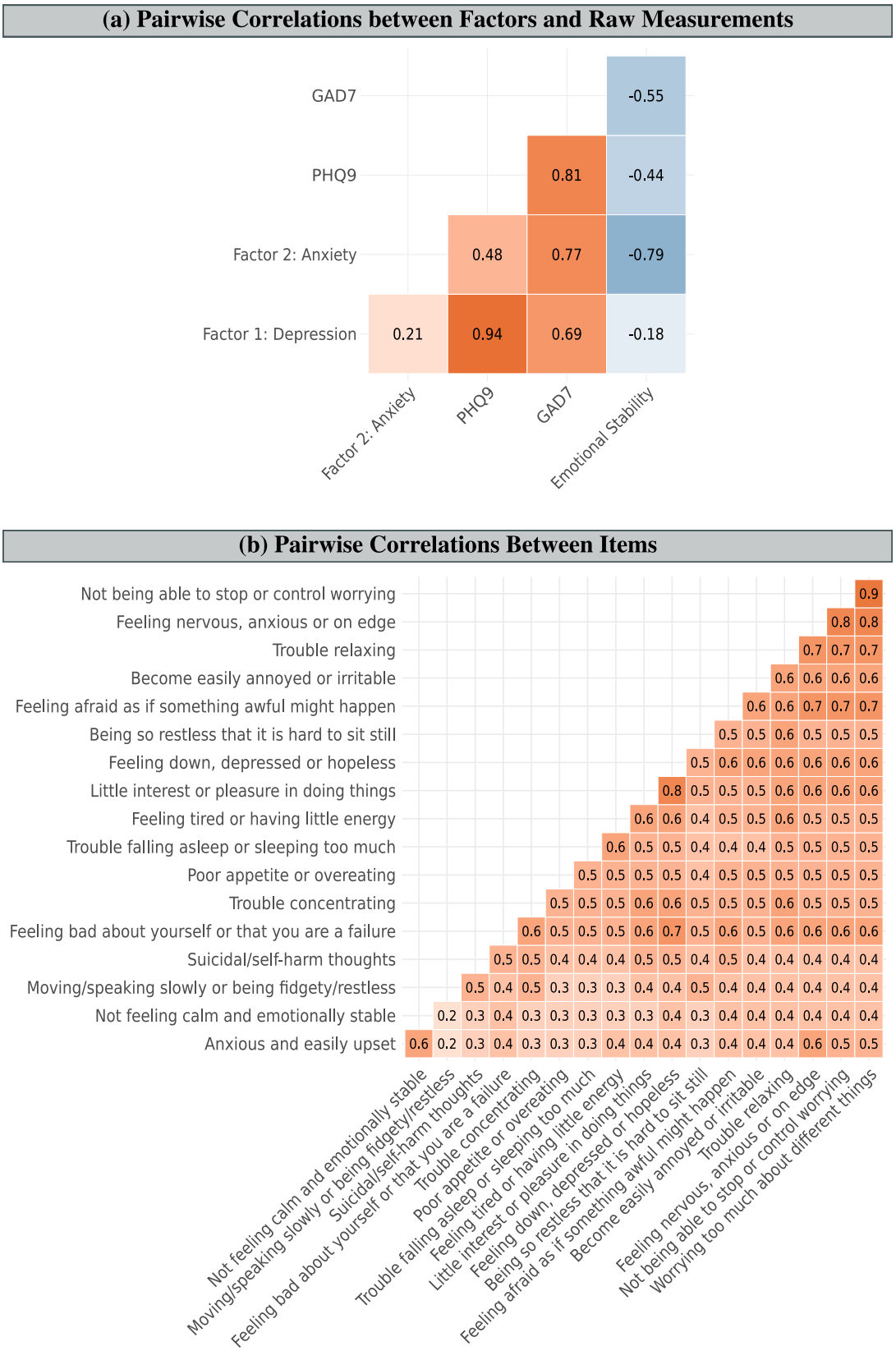


Figure 1: Pairwise Pearson correlations between factor scores and raw measures (a). Pairwise Pearson correlations between items in the factor analysis (b). Note that the names of some items are changed slightly to improve readability.

The three items "*Trouble relaxing*", "*Become easily annoyed or irritable*", and "*Feeling afraid as if something awful might happen*" were removed due to heavy cross-loadings.
The factor analysis was conducted using orthogonal varimax rotation which produced minimally correlated factors (Abdi, 2003). Principal axis factoring was used as the factor extraction method because the questionnaire data were not normally distributed (Costello & Osborne, 2019).

Factors were labeled by inspecting the factor loadings on each factor (see figure 2). The items with the highest loadings on the first factor were almost exclusively from the PHQ9, with the exception of the lowest scoring item: "*Being so restless that it is hard to sit still*". The first factor was therefore named *Depression* and accounted for 32% of the variance in the original data with a Chronbach's Alpha of $\alpha = .91$. The items with the highest loadings on the second factor included three items from the GAD7 and the two items from the TIPI measuring emotional stability. The second factor was therefore named *Anxiety* and accounted for 22% of the variance in the original data with a Chronbach's Alpha of $\alpha = .88$. Both scores of Chronbach's Alpha indicate a high level of internal consistency.

The bottom section of figure 1a shows the pairwise Pearson correlations of the factor scores and the raw measures of depression, anxiety, and emotional stability. The correlation between factor one and factor two was low because of the orthogonal rotation. Factor one was mostly correlated with PHQ9, and factor two was mostly correlated with GAD7 and emotional stability. However, GAD7 was also highly correlated with factor one, and PHQ9 was also moderately correlated with factor two. Similar results are found in Spitzer et al. (2006). Overall, the factor analysis was successful in reducing the dimensionality of the original three measures to two measures. It was also successful in producing measures that were less correlated than the raw measures, although not completely uncorrelated. The factor scores thus make it possible to assess depression and anxiety as more separate dimensions.

2.4 Patient groups

Patient groups in the factor scores were defined from the scores of the PHQ9 and GAD7. Patient groups in the PHQ9 and GAD7 are separated by cut off points which are chosen to maximize sensitivity and specificity: PHQ9 ≥ 10 and GAD7 ≥ 8 (Kroenke et al., 2001, 2007). Using these cut off points in the present study, 29% of participants were depressed and 26% of participants had anxiety. These percentiles were transferred to the factor scores such that participants with factor one scores above the 71st percentile were defined as depressed, and participants with factor two scores above the 74th percentile were defined as anxious. The patient groups were used to define target variables in the subsequent analyses. In the remaining paper, *depression* refers to factor one, and *anxiety* refers to factor two.

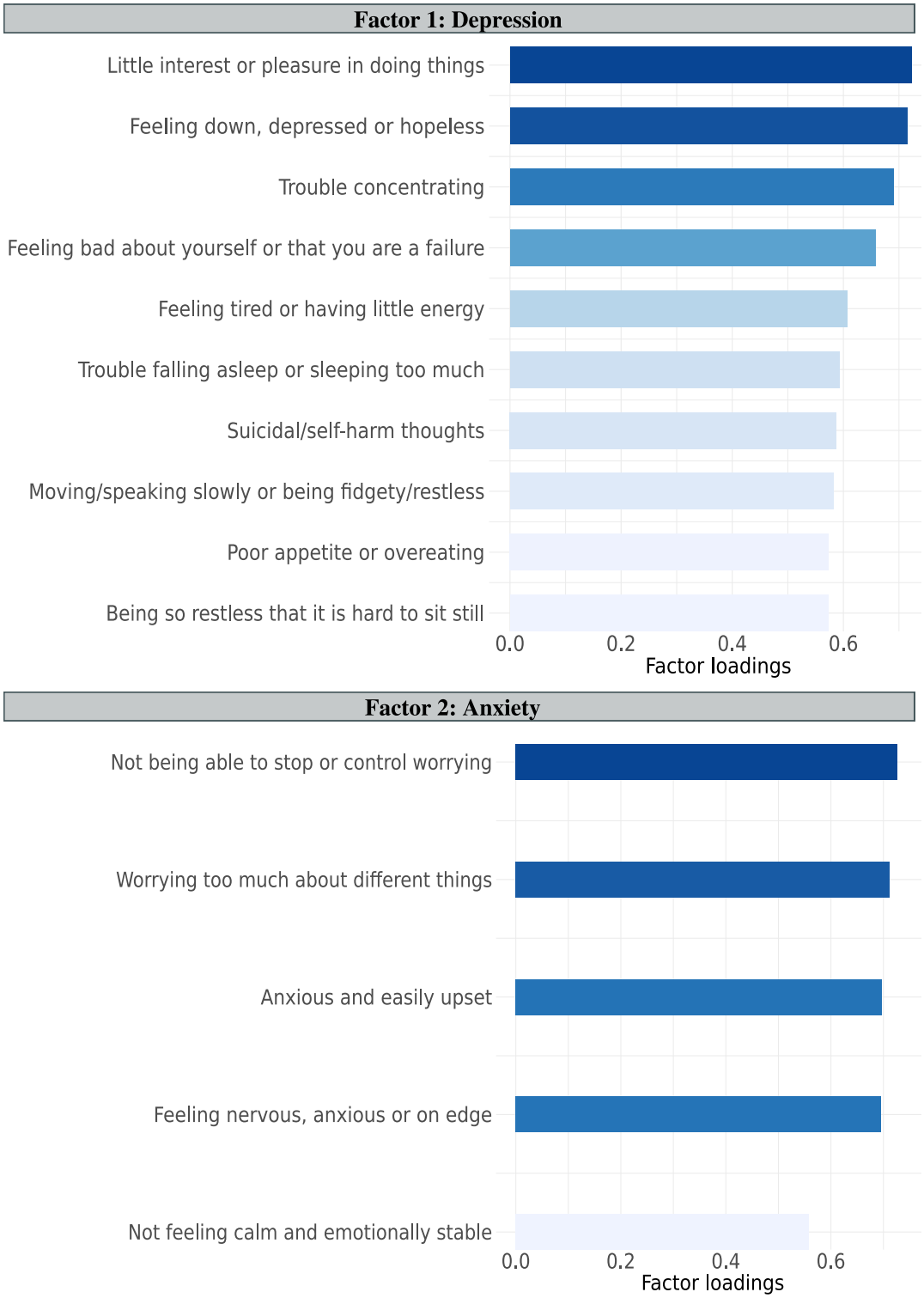


Figure 2: Item loadings on the two factors ordered by loading score. Note that the names of some items are changed slightly to improve readability.

2.5 Analysis 1: Mapping semantic landscapes of depression and anxiety

The first analysis investigated whether demonstratives could map differences in the semantic landscapes of patients and healthy individuals. To quantify the difference in demonstrative use, two scores were calculated for each of the 480 words, expressing the absolute difference in the proportion of proximal demonstratives between patients and healthy participants. The two scores are referred to as *Absolute Difference Proportion Proximal in Depression* (*ADPP-D*) and *Absolute Difference Proportion Proximal in Anxiety* (*ADPP-A*). A positive score means that the patient group used a larger proportion of the distal demonstrative, and a negative score means that the patient group used a larger proportion of the proximal demonstrative.

Each word was then assigned scores from eleven emotional dimensions. Three of the dimensions were from the VAD-lexicon: *Valence*, *Arousal*, and *Dominance* (S. Mohammad, 2018). The remaining eight dimensions were from the NRC-emotion-lexicon: *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, and *Trust* (S. M. Mohammad & Turney, 2013; S. Mohammad & Turney, 2010). Scores from the VAD-lexicon were continuous in the interval 0 to 1 (low to high) and scores from the NRC-emotion-lexicon were binary, either evoking the particular emotion (*evocative*) or not (*non-evocative*). Even though both lexicons contain many words (VAD: n=20,000, and NRC-emotion: n=2,000), there were 17 and 93 non-overlapping words for the VAD- and NRC-emotion-lexicons, respectively. The subsequent analysis was conducted separately for each lexicon in order to retain as much data as possible.

To examine which emotional dimensions (features) could predict ADPP-D and ADPP-A (targets), a machine learning pipeline was built using the library `scikit-learn v.0.24.2` (Pedregosa et al., 2011) in `Python v.3.6.6` (Van Rossum & Drake, 2009). The pipeline was repeated for each target and for each lexicon, resulting in four iterations.
The pipeline is depicted in the top section of figure 3. Each iteration began by splitting the dataset of words and dimensions into a train dataset (70%) and a holdout dataset (30%). The holdout dataset did not include any data from words that were part of the training set. The split was done with non-overlapping groups, which ensured no data leakage from the group structure from the train dataset to the holdout dataset. Then, seven linear regression models were created.
The first was a simple baseline regressor with no additional specifications.
The second regressor used a method of hyperparameter tuning called *grid search*. It searched over combinations of the following hyperparameters: regularization technique (lasso, ridge, and elastic net), the strength of the regularization, and the ratio of lasso/ridge in the elastic net. The combination of hyperparameters that maximized the cross-validation score was selected (Bergstra & Bengio, 2012).
The third regressor used a method of feature selection called *forward feature selection*. This method added features in a sequential fashion based on the cross-validation score. The subset of features that maximized the cross-validation score was selected and used for the remaining regressors (Ferri et al., 1994).
The fourth regressor used grid search to select the optimal hyperparameters similar to the second regressor, and was then passed onto the remaining regressors.
The fifth regressor used *Adaboost* to train multiple weak regressors sequentially, such that subsequent regressors focused more on difficult instances. Afterwards, the weak regressors were aggregated into one strong regressor (Freund et al., 1999).
The sixth regressor used *bagging* to construct a number of datasets, trained weak regressors on each dataset separately, and then used these to get an aggregated regressor (Breiman, 1996).
The seventh regressor used *early stopping* in a stochastic gradient descent algorithm, where a cross-validation score was calculated at each step of the gradient descent. Training of the regressor was terminated when the cross-validation score stopped improving (Prechelt, 2012).

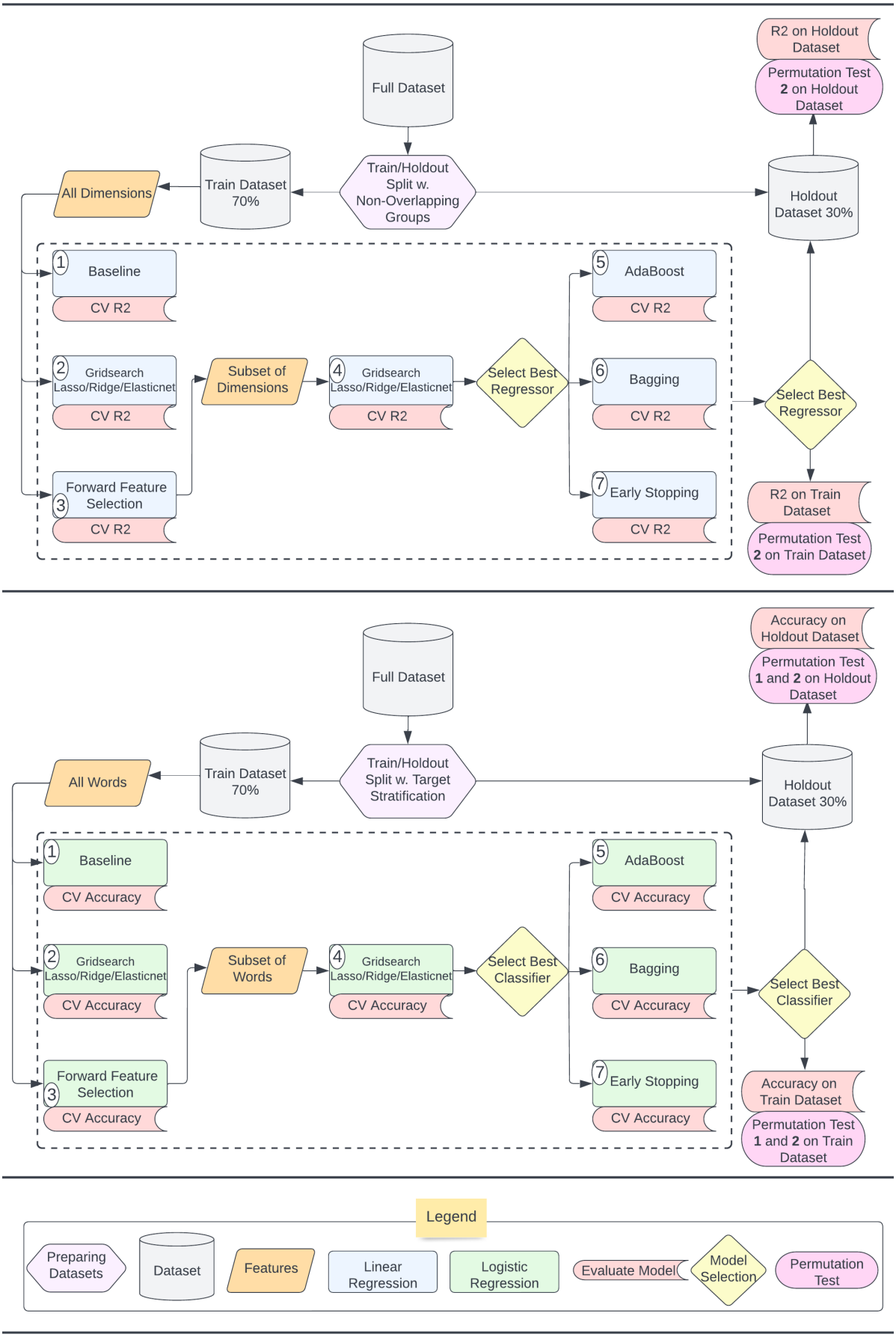


Figure 3: Machine learning pipeline for mapping semantic landscapes of depression and anxiety (top) and classifying depression and anxiety (bottom). CV = cross-validation.

All seven regressors were cross-validated with R^2 as the performance metric using *group-k-fold* with five folds. The regressor with the best cross-validated R^2 was selected and evaluated with R^2 on the train- and holdout datasets.

Lastly, permutation test 2 was conducted for the best regressor on the train- and holdout datasets. Permutation test 2 evaluated the importance of individual features, revealing which features were important for the performance of the regressor. Feature columns were permuted one at a time, thereby removing the dependency between the feature and the target, and a cross-validated R^2 was calculated. Feature importance was defined as the difference between the original R^2 and the R^2 from permuting the feature column. This process was repeated 100 times for each feature column, providing a mean feature importance and a standard deviation. Feature importance is thus a measure of how much the R^2 changed when a feature column was permuted. A feature was given high importance if the R^2 decreased substantially when the feature column was permuted. In contrast, a negative feature importance demonstrated that the feature was not important for the performance of the regressor (Ojala & Garriga, 2009).

2.6 Analysis 2: Training classifiers to predict depression and anxiety

The second analysis examined whether participants' response patterns in the DCT could predict depression and anxiety. Participants were divided into 8 groups in the experiment, and each group was presented with 60 words. Separate analyses were conducted for each group since they contain a unique subset of words.

To examine whether certain words (features) could predict depression and anxiety (targets) in the DTC, a machine learning pipeline was built and repeated for each group and each target, resulting in 16 iterations. As shown in the bottom section of figure 3, the overall pipeline was very similar to the pipeline from the previous analysis, nonetheless with some notable differences.
Each iteration began by splitting the dataset of participants' responses to each word into a train dataset (70%) and a holdout dataset (30%). The holdout dataset did not include any data from participants who were part of the training set. The split used target stratification, which ensured that the proportion of patients/non-patients was equal in train- and holdout datasets. The seven classifiers were created similarly to the previous pipeline and cross-validated using accuracy as the performance metric. All cross-validations used *repeated-stratified-k-fold* with five folds and five repetitions. The seven classifiers were logistic regression models. The best classifier was selected and evaluated with accuracy on the train- and holdout datasets.

Lastly, permutation tests 1 and 2 were conducted for the best classifier on the train- and holdout datasets. Permutation test 2 evaluated the importance of individual features based on accuracy. Permutation test 1 evaluated the overall performance of the classifier, by testing the probability of the observed cross-validated accuracy against the null. The null distribution was generated by permuting the target column and calculating the cross-validated accuracy with 1,000 repetitions. The null hypothesis thus assumed no dependency between features and target. A small p-value rejected the null hypothesis, indicating that the observed cross-validated accuracy was unlikely to be obtained by chance alone, meaning that the classifier exploited a real structure between features and target (Ojala & Garriga, 2009).

3 Results

3.1 Descriptive results

The overall proportion of proximal/distal demonstratives was 0.462/0.538.

3.2 Mapping semantic landscapes of depression and anxiety

Table 1 shows the features, performance (R^2) on the train- and holdout datasets, and also which regressor was selected for each lexicon and each target.
Regressor IV has three features. The pairwise partial Kendall correlations (τ) of these features were as follows: $\tau_{Anticipation:Disgust}$ = - .05, p = .54; $\tau_{Anticipation:Sadness}$ = - .06, p = .29; $\tau_{Disgust:Sadness}$ = .37, p < 0.001. Variance inflation factor (VIF) scores of the three features were: *Anticipation* = 1.00; *Disgust* = 1.27; *Sadness* = 1.28.

Lexicon	Nr.	Target	Features	R^2 Train	R^2 Holdout	Regressor Selected
VAD	I	ADPP-D	Valence	.075	.059	3
	II	ADPP-A	Valence	.051	.077	3
NRC Emotion	III	ADPP-D	Fear	.041	.016	3
	IV	ADPP-A	Anticipation, Sadness, Disgust	.096	.052	3

Table 1: Features, performance (R^2) on train- and holdout datasets, and the regressor selected for each lexicon and each target.

Figure 4 shows the mean feature importance and standard deviation for the emotional dimensions in train- and holdout datasets.

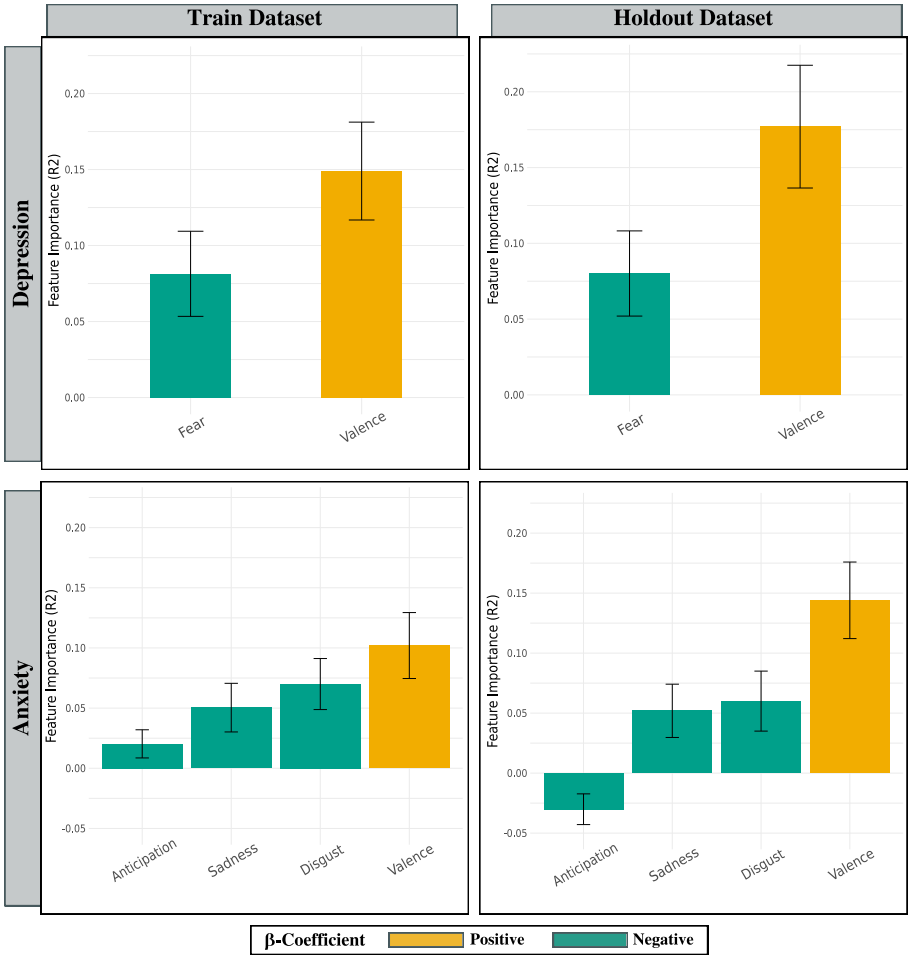


Figure 4: Mean feature importance and standard deviation of the features in the regressors. Green bars denote a negative β-coefficient; higher values of the feature yielded a larger proportion of *this* in the patient group. Yellow bars denote a positive β-coefficient; higher values of the feature yielded a larger proportion of *that* in the patient group.

3.3 Classifying depression and anxiety

Figure 5 shows the performance of classifiers in permutation test 1. The distributions show the permuted cross-validated accuracy, and the dotted lines show the observed cross-validated accuracy and its significance level. A red line indicates a significant accuracy; a blue line indicates a non-significant accuracy.

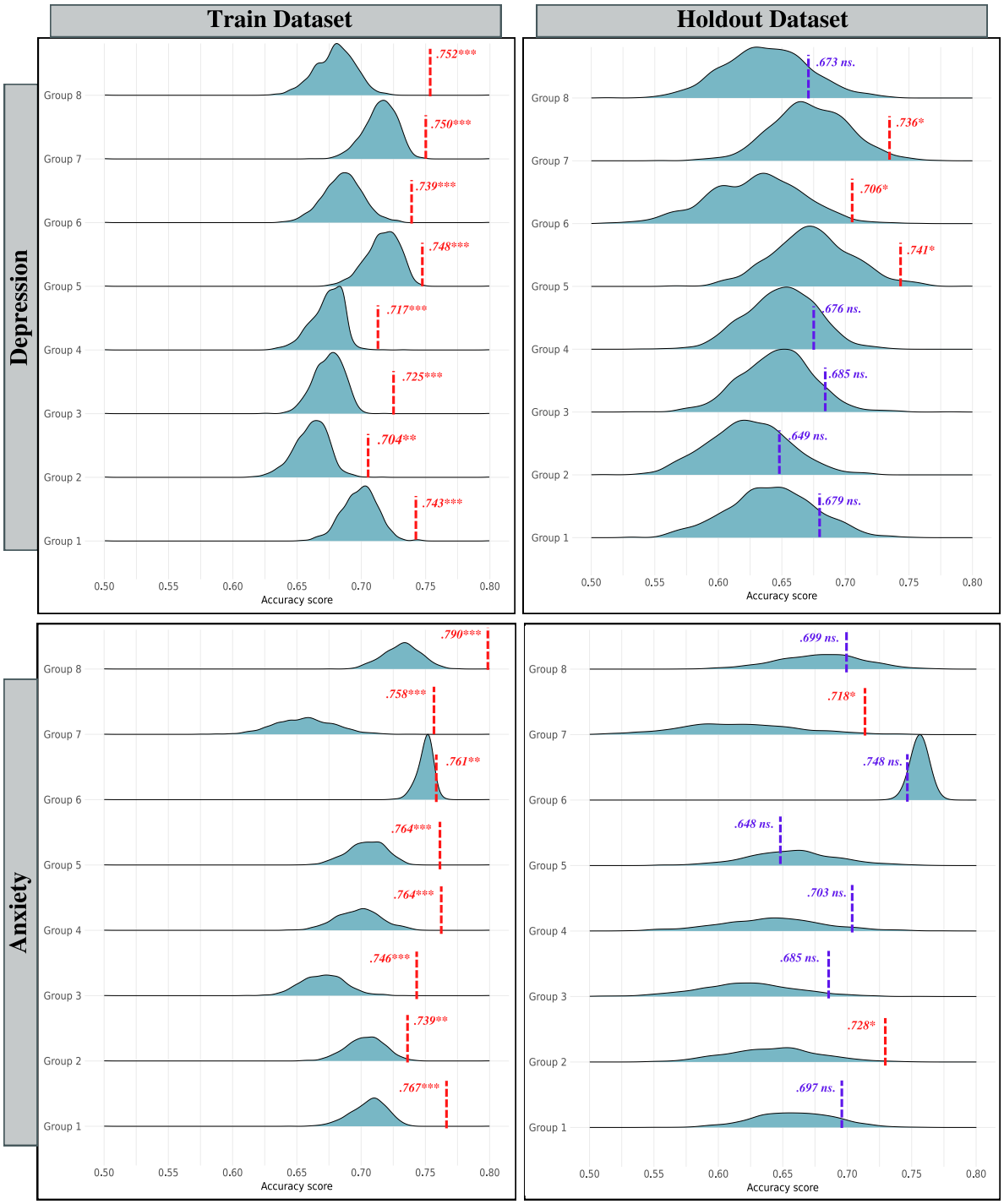


Figure 5: Performance of classifiers in permutation test 1. Distributions show the permuted cross-validated accuracy (number of repetitions = 1000). The dotted lines show the observed cross-validated accuracy and its significance level. A red line indicates a significant accuracy; a blue line indicates a non-significant accuracy. $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p > 0.05$ (ns.).

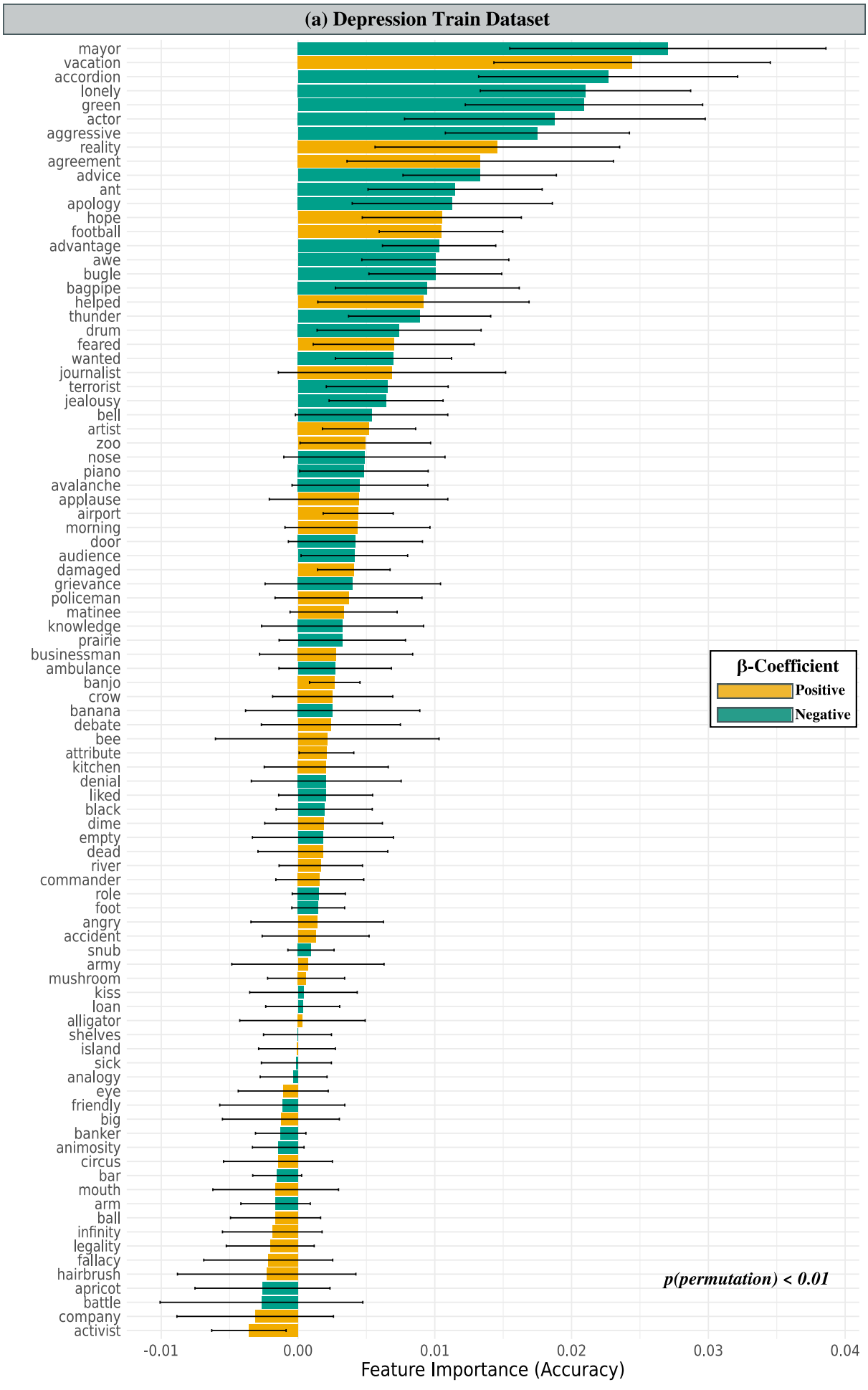
Table 2 expands the results from figure 5, showing the proportion of non-patients, the mode of the permutation distributions, and the observed accuracy on the train- and holdout datasets. It also shows which classifier was selected.

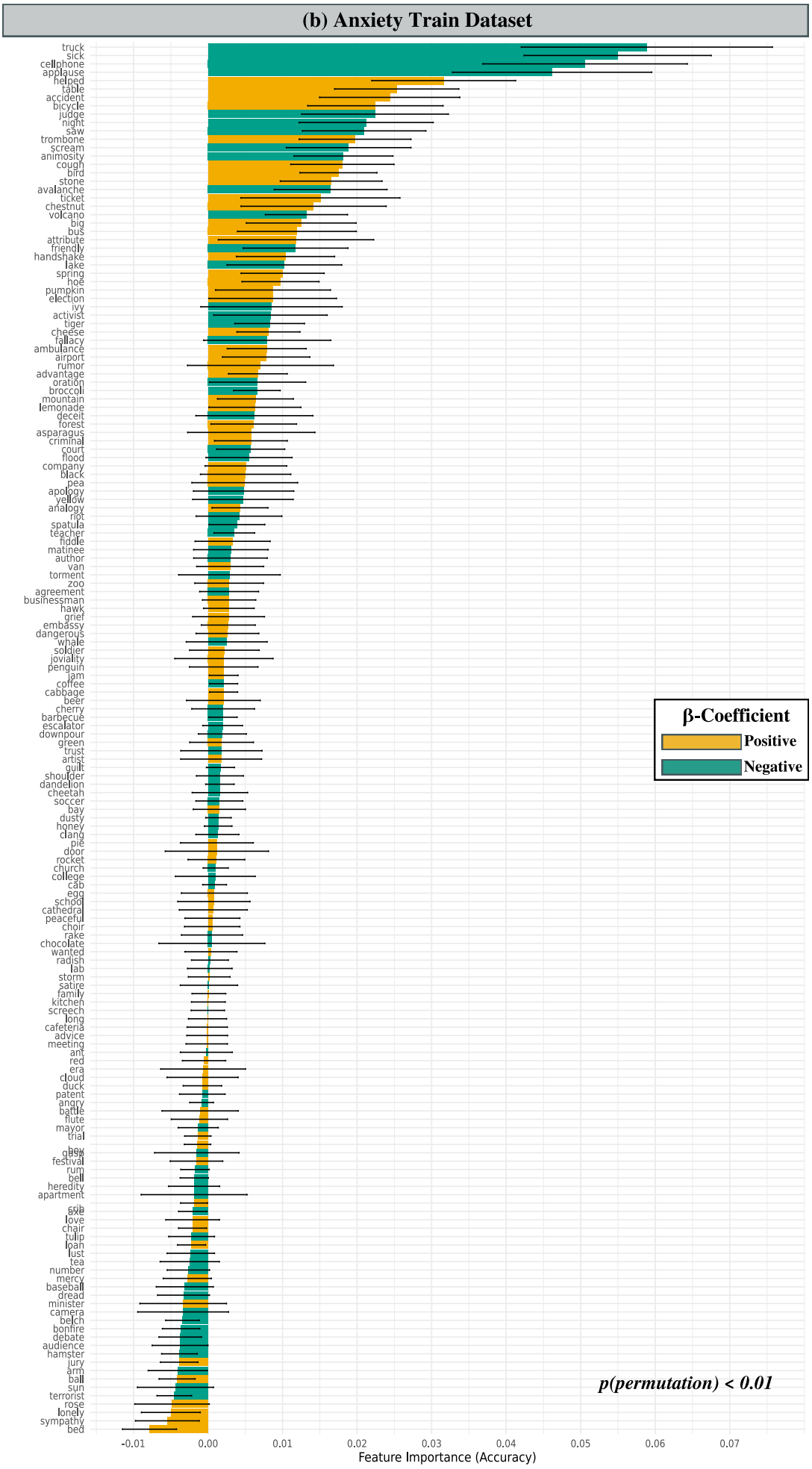
Target	Group	Prop. Non-Patient Train	Prop. Non-Patient Holdout	Permutation Mode Train	Permutation Mode Holdout	Accuracy Train	Accuracy Holdout	Classifier Selected
Depression	1	.723	.716	.704	.633	*.743* ***	.679 ns.	3
	2	.678	.675	.670	.623	*.704***	.649 ns.	3
	3	.690	.694	.690	.661	*.725****	.685 ns.	3
	4	.686	.694	.686	.657	*.717****	.676 ns.	3
	5	.736	.741	.724	.676	*.748****	*.741**	3
	6	.721	.714	.688	.630	*.739****	*.706**	3
	7	.734	.736	.727	.664	*.750****	*.736**	3
	8	.706	.708	.679	.647	*.752****	.673 ns.	3
Anxiety	1	.727	.725	.712	.670	*.767****	.697 ns.	3
	2	.735	.728	.708	.658	*.739***	*.728**	3
	3	.721	.726	.676	.629	*.746****	.685 ns.	3
	4	.752	.748	.705	.640	*.764****	.703 ns.	4†
	5	.736	.741	.716	.667	*.764****	.648 ns.	3
	6	.754	.756	.754	.756	*.761***	.748 ns.	3
	7	.734	.736	.660	.591	*.758****	*.718**	4†
	8	.763	.763	.737	.681	*.790****	.699 ns.	3

Table 2: Proportion of non-patients, the mode of permutation scores, and the observed accuracy on train- and holdout datasets, and which classifier was selected. †: The selected hyperparameters of the grid search were: Elastic net, inverse of regularization strength $C = 1$, and lasso/ridge-ratio = 0.1. For exact implementation, see the documentation on the scikit-learn webpage.

Figure 6 on pages 16, 17, and 18 shows the mean feature importance and standard deviation for the words in the classifiers that had a significant p-value in permutation test 1 for both train- and holdout datasets.

The pairwise partial Kendall correlations of the features in the classifiers were between -0.2 to 0.3 (see Appendix B for all pairwise correlations). Variance inflation factor (VIF) scores of features in the classifiers were between 1.01 to 1.31 (see Appendix C for all VIF scores).


(a) Depression Train Dataset
mayor
vacation
accordion
lonely
green
actor
aggressive
reality
agreement
advice
ant
apology
hope
football
advantage
awe
bugle
bagpipe
helped
thunder
drum
feared
wanted
journalist
terrorist
jealousy
bell
artist
zoo
nose
piano
avalanche
applause
airport
morning
door
audience
damaged
grievance
policeman
matinee
knowledge
prairie
businessman
ambulance
banjo
crow
banana
debate
bee
attribute
kitchen
denial
liked
black
dime
empty
dead
river
commander
role
foot
angry
accident
snub
army
mushroom
kiss
loan
alligator
shelves
island
sick
analogy
eye
friendly
big
banker
animosity
circus
bar
mouth
arm
ball
infinity
legality
fallacy
hairbrush
apricot
battle
company
activist
β-Coefficient
Positive
Negative
p(permutation) < 0.01
-0.01
0.00
0.01
0.02
0.03
0.04
Feature Importance (Accuracy)


(b) Anxiety Train Dataset
truck
sick
cellphone
applause
helped
table
accident
bicycle
judge
night
saw
trombone
scream
animosity
cough
bird
stone
avalanche
ticket
chestnut
volcano
big
bus
attribute
friendly
handshake
lake
spring
hoe
pumpkin
election
ivy
activist
tiger
cheese
fallacy
ambulance
airport
rumor
advantage
oration
broccoli
mountain
lemonade
deceit
forest
asparagus
criminal
court
flood
company
black
pea
apology
yellow
analogy
riot
spatula
teacher
fiddle
matinee
author
van
torment
zoo
agreement
businessman
hawk
grief
embassy
dangerous
whale
soldier
jovality
penguin
jam
coffee
cabbage
beer
cherry
barbecue
escalator
downpour
green
trust
artist
guilt
shoulder
dandelion
cheetah
soccer
bay
dusty
honey
clang
pie
door
rocket
church
college
cab
egg
school
cathedral
peaceful
choir
rake
chocolate
wanted
radish
lab
storm
satire
family
kitchen
screech
long
cafeteria
advice
meeting
ant
red
era
cloud
duck
patent
angry
battle
flute
mayor
trial
boy
gasp
festival
rum
bell
heredity
apartment
cube
axe
love
chair
tulip
loan
lust
tea
number
mercy
baseball
dread
minister
camera
belch
bonfire
debate
audience
hamster
jury
arm
ball
sun
terrorist
rose
lonely
sympathy
bed
β-Coefficient
Positive
Negative
p(permutation) < 0.01
-0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Feature Importance (Accuracy)

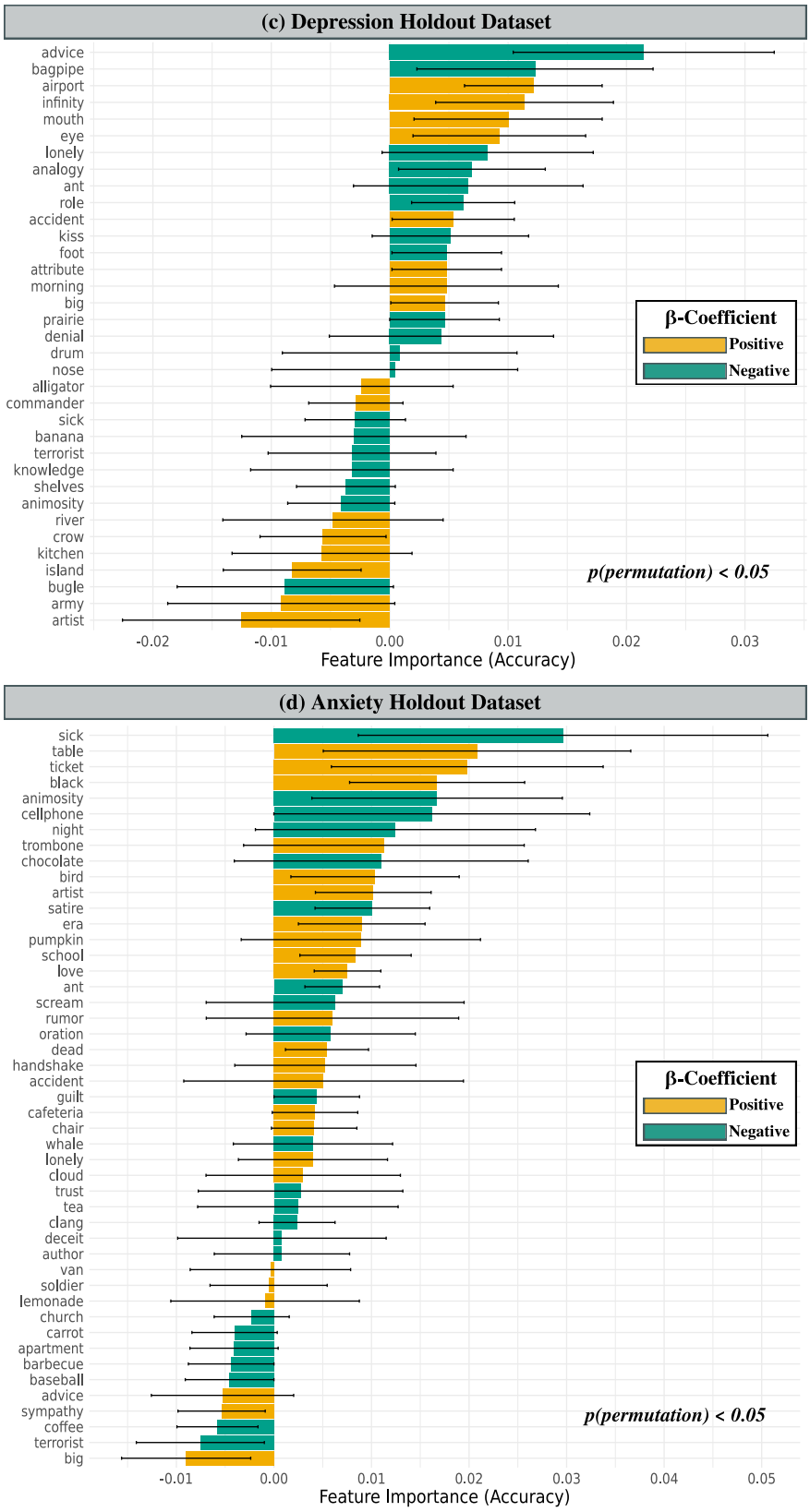


Figure 6: Mean feature importance and standard deviation of features in the classifiers that had a significant p-value in permutation test 1. Each subfigure is a combination of the target (depression or anxiety) and the dataset (train or holdout). Green bars denote a negative β-coefficient; the odds of choosing *this* was higher in the patient group. Yellow bars denote a positive β-coefficient; the odds of choosing *that* was higher in the patient group. Note that the feature importance of individual words reflects the accuracy of its classifier, and since all classifiers had a slightly different accuracy, the feature importances of words from different classifiers are not directly comparable.

4 Discussion

4.1 Mapping semantic landscapes of depression and anxiety

The aim of the first analysis was to investigate whether demonstratives could map the differences between the semantic landscapes of participants with depression or anxiety and healthy participants. The results demonstrate that the difference in the absolute proportion of proximal demonstratives between depressed and non-depressed participants was predicted by *Valence* and *Fear* (see table 1). Moreover, the standard deviation for the feature importance of *Valence* and *Fear* did not cross zero in either train- or holdout datasets, demonstrating that *Valence* and *Fear* contributed substantially to the prediction power of the regressors (see the top section of figure 4). The nine remaining dimensions - *Arousal*, *Dominance*, *Anger*, *Anticipation*, *Disgust*, *Joy*, *Sadness*, *Surprise*, and *Trust* - were discarded in the forward feature selection procedure.

Furthermore, the difference in the absolute proportion of proximal demonstratives between anxious and non-anxious participants was predicted by *Valence*, *Anticipation*, *Sadness*, and *Disgust* (see table 1). The seven remaining dimensions – *Arousal*, *Dominance*, *Anger*, *Fear*, *Joy*, *Surprise*, and *Trust* - were discarded in the procedure of forward feature selection. Moreover, the standard deviation for the feature importance of *Valence*, *Sadness*, and *Disgust* did not cross zero in either train- or holdout datasets (see the bottom section of figure 4). In contrast, the feature importance of *Anticipation* on the holdout dataset was negative, indicating that *Anticipation* did not contribute to the prediction power of the regressor. *Anticipation* was therefore discarded as a predictive feature.
Regarding the multicollinearity of *Sadness* and *Disgust* in regressor IV (see table 1), it was found that their VIF scores were below the often-used cut off point of five, indicating no issue with multicollinearity despite their significant moderate correlation (Sheather, 2009, p.213).

The performance (R^2) of regressors I, II, and III generalized to the holdout dataset, although dropping compared to the train dataset, indicating that these regressors did not suffer from major overfitting (see table 1). In contrast, regressor IV unexpectedly performed better on the holdout dataset (see table 1). This is surprising since none of the holdout data was used for training the regressor, and the possibility of data leakage from the group structure was accounted for in the train/holdout split. One explanation could be that the holdout data was more agreeable by chance, meaning that the relationship between the target and the features was more pronounced. If this was the cause, some solutions could be to make a new train/holdout split, and make sure that the distributions of the features and target were similar. It could also help to increase the number of words in the experiment to avoid unbalanced train/holdout splits.
Given the modest R^2 and generalizability of the regressors to the holdout dataset, it is argued that the DCT successfully mapped differences in the semantic landscapes of participants with depression or anxiety and healthy participants. Nonetheless, the R^2 of all four regressors were relatively modest, especially in the holdout data, meaning that most of the variance in demonstrative use between patients and non-patients remains unexplained/unpredicted by the emotional dimensions.

Following the hypothesis from Rocca & Wallentin (2020) that the proximal/distal contrast in demonstrative use relates to the distance between the speaker and the referent in a semantic space, this study demonstrated that certain emotional dimensions elicited more proximal demonstratives in participants with depression and anxiety whereas other dimensions elicited more distal demonstratives. For participants with depression, words rated as more positive yielded a larger proportion of distal demonstratives, whereas words that evoked fear yielded a larger proportion of proximal demonstratives (see figure 4). These results suggest that depressed

individuals navigate a semantic landscape where negative and fearful referents are considered closer to the self than positive and harmless referents. Moreover, for participants with anxiety, words rated as more positive also yielded a larger proportion of distal demonstratives, whereas words that evoked sadness and disgust yielded a larger proportion of proximal demonstratives (see figure 4). These results suggest that anxious individuals experience a semantic landscape where negative, sad, and disgusting referents are considered closer to the self. The characteristics of the semantic landscapes are not surprising given the symptoms of depression and anxiety (Tiller, 2013).
In general, the results suggest that the organization of the multidimensional semantic space is altered in individuals with depression and anxiety. Patients interpret their selves as being situated in a world that is negative, fearful, sad, and disgusting, and where the brighter aspects of existence seem further away from themselves.

4.2 Classifying depression and anxiety

The second analysis aimed to examine whether participants' response patterns in the DCT could predict depression and anxiety. It was demonstrated that all classifiers had better accuracy than chance in the train dataset (see table 2). In contrast, on the holdout dataset, only three (groups 5, 6, and 7) and two (groups 2 and 7) classifiers had a significant accuracy for depression and anxiety, respectively (see table 2). The failure to predict new instances in some groups is a symptom of overfitting, which happened despite the procedure of cross-validation and regularization. Despite the drop in performance on the holdout dataset, it is argued that the classifiers were overall successful at predicting depression and anxiety.

The classifiers were evaluated on how much their performance metric deviated from the permuted distributions. To better understand the performance of the classifiers, a closer look at the permuted distributions is therefore needed.
Since accuracy was the performance metric, the mode of the distributions should be affected by the proportion of non-patients in the group. However, the proportion of non-patients was not entirely the same in the train- and holdout datasets because the stratification in the train/holdout split was conducted with respect to depression and anxiety *at the same time*, meaning that the proportion of non-patients had to be balanced with respect to two parameters (see table 2). Therefore, some of the variations in the permuted distributions are caused by the train/holdout splitting procedure. Moreover, there was a greater variation in the holdout distributions because the holdout datasets only contained 30% of the data. Lastly, there was the following hierarchy across all classifiers (see table 2):

$$proportion\,non\,patient \geq permutation\,mode\,train\,dataset \geq permutation\,mode\,holdout\,dataset$$

This hierarchy exists because cross-validation was applied in the calculation of the permuted distributions. The modes of the train permutations were lower than the proportion of non-patients because the permuted classifier learned incorrect dependencies between the features and target in the train dataset, thereby predicting poorly in the validation dataset. Moreover, the permutation modes of the holdout datasets were lower than the permutation modes of the train datasets because classifiers were predicting on completely novel instances in the holdout dataset.

The second aim of the analysis was to examine which words were useful for classifying depression and anxiety in the DCT. Appendix A shows the full list of words for each group. Words that were included in classifiers which were significant in permutation test 1 on the train dataset are coloured light brown. Words that were included in classifiers which were significant in permutation test 1 on the train- *and* holdout datasets are coloured dark brown. Not all words are coloured since the features in all classifiers were only a subset of words as chosen by the

procedure of forward feature selection and elastic net (see table 2).
In addition, figure 6 shows the mean feature importance of words from classifiers that were significant in permutation test 1. The standard deviation of the majority of the words, especially on the holdout dataset, overlap with zero, meaning that these words did not contribute to the performance of their respective classifier. Nonetheless, a subset of words consistently improved their classifiers' performances across all 100 permutations, even on the holdout dataset, suggesting that these specific words were key for classifying depression and anxiety. While most of these words are difficult to relate to the semantic landscape of depressed and anxious individuals, there are some notable exceptions. Patients had *lower* odds of choosing the distal demonstrative for *sick* and *lonely* and *higher* odds of choosing the distal demonstrative for *helped* and *school*.
It is speculated that the words in the DCT are the driving factor that determines whether demonstratives can probe individual differences in personality. However, the words in the current study were chosen with the purpose of conducting an analysis with 76 cognitively relevant semantic dimensions (Binder et al., 2016; Lynott et al., 2020) similar to Rocca & Wallentin (2020), meaning that the words were *not* chosen for classifying depression and anxiety. Therefore, the groups whose classifiers were significant on the holdout dataset were successful because they consisted of performance-driving words by coincidence.

On a last note, there was no tendency of multicollinearity among the words in the classifiers. As shown in appendix B, the maximum value of the pairwise partial Kendall correlations was 0.3. Moreover, as shown in appendix C, the maximum VIF score was 1.31, which is below the often-used cut off point of five (Sheather, 2009, p. 213).

4.3 Implications, limitations, and future directions

Previous research argued that demonstrative choice is modulated by a range of factors common among individuals, including physical distance, visibility, ownership, familiarity (Coventry et al., 2014), social context (Rocca, Wallentin, et al., 2019), manipulability, valence and the self (Rocca, Tylén, et al., 2019; Rocca & Wallentin, 2020).
This study found that demonstratives were successful at mapping individual differences in semantic landscapes and classifying depression and anxiety. This provides evidence that demonstrative choice is affected, not only by factors common across individuals, but also by the way experiences, memories, mental health, personality traits, etc. shape the individual semantic landscape.

The next step in using the DCT to examine individual differences is to carefully select a set of words tailored to the personality dimension of interest. Such a set of tailored words is expected to increase the predictive power of the models. Future studies whose purpose is to classify depression and anxiety could select words that were driving the performance of the classifiers in this study (see figure 6). They could also select words according to the emotional dimensions found in this study (see figure 4). Moreover, future studies also need to investigate whether demonstratives can map other dimensions of personality, such as the Big Five personality traits.

On a technical note, neither Adaboost, bagging or early stopping were selected as the best regressors or classifiers in the analysis (see table 1 and 2). This was unexpected, because the three models were meant to deal with overfitting. Future studies could improve their performance by increasing the amount of data and more carefully adjust their respective hyperparameters. Moreover, future studies could also utilize other machine learning models such as support vector machines, nearest neighbors algorithms, and decision trees.

As a final note, the DCT only examines the relationship between demonstratives and semantics in isolation, excluding the factors that are present in exophoric contexts. However, it is likely that physical and social factors interact with the semantic space of the individual to dynamically shape the use of demonstratives. Further studies need to investigate if and how these factors influence each other in a naturalistic setting.

5 Conclusion

This study found that the DCT was successful at mapping individual differences in semantic landscapes and classifying depression and anxiety, providing evidence that demonstrative choice is influenced by individual factors. The same referent can elicit different responses in individuals, making demonstratives incredibly versatile and flexible. Moreover, the results suggest that the organization of semantic space is altered in individuals with depression and anxiety. Demonstratives can thus reveal core aspects of individual semantic knowledge, underpinning their role as a powerful interface between language, cognition, and mental illnesses. Carefully selecting the set of words in future studies may increase the predictive power of the DCT.

References

Abdi, H. (2003). Factor rotations in factor analyses. , 792–795.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. (2).

Berti, A., & Frassinetti, F. (2000). When far becomes near: Remapping of space by tool use. (3), 415–420. doi: 10.1162/089892900562237

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. (3), 130–174. doi: 10.1080/02643294.2016.1147426

Breiman, L. (1996). Bagging predictors. (2), 123–140. doi: 10.1007/BF00058655

Bühler, K. (1934). *Sprachtheorie*. Jena Fischer.

Caldano, M., & Coventry, K. R. (2019). Spatial demonstratives and perceptual space: To reach or not to reach? doi: 10.1016/j.cognition.2019.06.001

Clark, E. V., & Sengul, C. J. (1978). Strategies in the acquisition of deixis*. (3), 457–475. doi: 10.1017/S0305000900002099

Cooperrider, K. (2016). The co-organization of demonstratives and pointing gestures. (8), 632–656. doi: 10.1080/0163853X.2015.1094280

Costello, A., & Osborne, J. (2019). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. (1). doi: https://doi.org/10.7275/jyj1-4868

Coventry, K. R., Griffiths, D., & Hamilton, C. J. (2014). Spatial demonstratives and perceptual space: Describing and remembering object location. , 46–70. doi: 10.1016/j.cogpsych.2013.12.001

Coventry, K. R., Guijarro-Fuentes, P., & Valdés, B. (2008). Second language acquisition of spatial terms in english and spanish.

Coventry, K. R., Valdés, B., Castillo, A., & Guijarro-Fuentes, P. (2008). Language within your reach: Near–far perceptual space and spatial demonstratives. (3), 889–895. doi: 10.1016/j.cognition.2008.06.010

Diessel, H. (2006). Demonstratives, joint attention, and the emergence of grammar. (4), 463–489. doi: 10.1515/COG.2006.015

Diessel, H. (2013). Where does language come from? some reflections on the role of deictic gesture and demonstratives in the evolution of language. (2), 239–249. doi: 10.1515/langcog-2013-0017

Diessel, H. (2014). Demonstratives, frames of reference, and semantic universals of space. (3), 116–132. doi: 10.1111/lnc3.12066

Diessel, H., & Coventry, K. R. (2020). Demonstratives in spatial language and social interaction: An interdisciplinary review.

Diessel (2005). Distance Contrasts in demonstratives. In Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of language structures (pp. 170-173)*. OUP Oxford.

Ellis, L., Hoskin, A. W., & Ratnasingam, M. (2018). *Handbook of social status correlates*. Academic Press.

Farnè, A., Bonifazi, S., & Làdavas, E. (2005). The role played by tool-use and tool-length on the plastic elongation of peri-hand space: a single case study. (3), 408–418. doi: 10.1080/02643290442000112

Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. In E. S. Gelsema & L. S. Kanal (Eds.), *Machine intelligence and pattern recognition* (pp. 403–413). doi: 10.1016/B978-0-444-81892-8.50040-7

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. (771), 1612.

Geoffryy, L. (2016). *Word frequencies in written an spoken english*. Longman.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. (6), 504–528. doi: 10.1016/S0092-6566(03)00046-1

Gudde, H. B., Coventry, K. R., & Engelhardt, P. E. (2016). Language and memory for object location. , 99–107. doi: 10.1016/j.cognition.2016.04.016

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. (2), 179–185. doi: 10.1007/BF02289447

Kroenke, K., Spitzer, R. L., Williams, J. B., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. (5), 317–325.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. (9), 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Longo, M. R., & Lourenco, S. F. (2006). On the nature of near space: Effects of tool use and the transition to far space. (6), 977–981. doi: 10.1016/j.neuropsychologia.2005.09.003

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. (3), 1271–1291. doi: 10.3758/s13428-019-01316-z

Maes, A. (2007). Spatial and conceptual demonstratives. , 127–144.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 174–184). Association for Computational Linguistics. doi: 10.18653/v1/P18-1017

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. (3), 436–465. doi: 10.1111/j.1467-8640.2012.00460.x

Ojala, M., & Garriga, G. C. (2009). Permutation tests for studying classifier performance. In *2009 ninth IEEE international conference on data mining* (pp. 908–913). (ISSN: 2374-8486) doi: 10.1109/ICDM.2009.108

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in python. , 2825–2830.

Peeters, D., Krahmer, E., & Maes, A. (2021). A conceptual framework for the study of demonstrative reference. (2), 409–433. doi: 10.3758/s13423-020-01822-8

Plutchik, R. (1994). *The psychology and biology of emotion.* HarperCollins College Publishers.

Prechelt, L. (2012). Early stopping — but when? In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade: Second edition* (pp. 53–67). Springer. doi: 10.1007/978-3-642-35289-8_5

R Core Team. (2021). R: A language and environment for statistical computing.

Revelle, W. (2022). psych: Procedures for psychological, psychometric, and personality research. (R package version 2.2.9)

Rocca, R., Tylén, K., & Wallentin, M. (2019). This shoe, that tiger: Semantic properties reflecting manual affordances of the referent modulate demonstrative use. (1). doi: 10.1371/journal.pone.0210333

Rocca, R., & Wallentin, M. (2020). Demonstrative reference and semantic space: A large-scale demonstrative choice task study.

Rocca, R., Wallentin, M., Vesper, C., & Tylén, K. (2019). This is for you: Social modulations of proximal vs. distal space in collaborative interaction. (1), 14967. doi: 10.1038/s41598-019-51134-8

Russell, J. A. (2003). Core affect and the psychological construction of emotion. , 145–172. doi: 10.1037/0033-295X.110.1.145

Sheather, S. (2009). *A modern approach to regression with r.* Springer. doi: 10.1007/978-0-387-09608-7

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. (10), 1092–1097. doi: 10.1001/archinte.166.10.1092

Tiller, J. W. G. (2013). Depression and anxiety. (6).

Todisco, E., Rocca, R., & Wallentin, M. (2021). The semantics of spatial demonstratives in spanish: a demonstrative choice task study. (4), 503–533. doi: 10.1017/langcog.2021.11

Van Rossum, G., & Drake, F. L. (n.d.). *Python 3 reference manual*. CreateSpace.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. (3). (Number: 3) doi: 10.33151/ajp.8.3.93

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. (6), 1100–1122. doi: 10.1177/1745691617693393

Appendix

A

Full list of words

Group 1 Depression	Group 2 Depression	Group 3 Depression	Group 4 Depression	Group 5 Depression	Group 6 Depression	Group 7 Depression	Group 8 Depression
Word	Word	Word	Word	Word	Word	Word	Word
helped	aggressive	activist	accordion	damaged	alligator	accident	actor
advantage	angry	ambulance	apology	airport	army	advice	agreement
applause	barbecue	avalanche	apricot	analogy	bagpipe	animosity	asparagus
awe	bell	bonfire	bay	arm	banana	ant	axe
banjo	book	broccoli	beer	attribute	bar	apartment	ball
banker	business	businessman	cathedral	audience	bed	artist	battle
belief	cab	camel	cheetah	bus	blueberry	author	beach
bicycle	cabbage	camera	college	car	boy	baseball	bee
blue	cafeteria	cherry	corkscrew	carriage	bribe	big	belch
bread	carrot	choir	dandelion	chestnut	butterfly	bird	child
bridge	chair	cranberry	dangerous	chicken	carnival	black	company
cabinet	church	criminal	delirium	computer	cash	boat	couple
cheese	clang	curse	dictation	corn	cold	bugle	cyclone
circus	clarinet	dusty	embassy	court	commander	cellphone	debate
clever	coffee	election	empty	desk	crib	chime	dime
clue	dead	escalator	explosion	egg	crow	chipmunk	dinner
cough	doctor	evening	fence	eye	denial	chocolate	diplomat
council	eggplant	expensive	fiddle	festival	dog	cloud	dolphin
cucumber	embrace	friendly	fish	folly	downpour	comb	door
dark	envy	fun	flower	foot	dread	deceit	elevator
day	etiquette	garden	flute	funeral	drum	elephant	fallacy
driver	excuse	green	fountain	girl	duck	era	famous
farmer	farm	grief	gasp	gum	editor	fee	feather
faucet	finger	grievance	goldfish	hailstorm	elm	gunshot	flood
feared	fireworks	happy	hall	ham	engineer	handshake	gratitude
field	football	harp	honey	heavy	family	heredity	hairbrush
forest	gong	highway	insult	heroism	fate	keyboard	harmonica
glass	guard	hot	ire	hierarchy	hamster	lemonade	injured
hand	guilt	hygiene	jam	horse	hawk	lonely	intellect
hoe	hair	jaw	lab	hurricane	island	love	ivy
honeymoon	home	joviality	landslide	infinity	jungle	medicine	journalist
hope	jealousy	judge	leg	irony	jury	motive	law
hospital	lawyer	limousine	man	joke	ketchup	mouse	lightning
hotel	magazine	loud	mandolin	joy	kiss	muscle	liked
ice	moose	luck	matinee	key	kitchen	mustard	lip
lake	moral	malice	patent	lived	knowledge	newspaper	loan
mushroom	musical	mayor	pea	majority	legality	night	monkey
old	new	megaphone	pilot	mob	long	oration	mystery
park	paradox	mercy	plot	mountain	lust	peace	office
patient	pen	minister	protest	parent	meeting	pumpkin	optimism
piano	pineapple	noun	reality	power	morning	rumor	pan
pig	plane	oak	red	prairie	mosquito	satire	party
plum	plea	parade	riot	prison	mouth	school	pencil
reporter	priest	pie	sandpaper	role	nose	scream	penguin
restaurant	raspberry	policeman	scientist	scissors	number	sick	perjury
ricochet	sailboat	rake	shoulder	shiny	peaceful	soldier	powerful
semester	saxophone	rocket	soccer	small	politician	sun	problem
sin	screech	rose	speech	snake	radish	sympathy	quantity
snub	shame	salmon	stapler	squeal	river	symphony	rally
spring	tangerine	saw	student	stampede	shelves	table	rum
teacher	turtle	street	submarine	stone	soft	tax	scooter
thunder	verb	theme	subway	sum	spaghetti	tea	sled
tiger	water	tobacco	theater	summer	spatula	terrorist	store
tomato	wealthy	toe	theory	team	spiritual	ticket	storm
truce	winter	tornado	torment	television	testimony	tribute	toaster
trumpet	witness	tourist	voter	train	tired	trombone	tree
volcano	woman	treaty	wanted	truck	trial	trust	truth
white	year	vacation	window	used	tulip	tuba	victim
whole	young	vice	worker	whine	umbrella	van	woe
xylophone	zone	worth	zoo	yellow	wit	whale	wonder

Figure A1: Full list of words for each group. Words that were included in **depression**-classifiers that were significant in permutation test 1 on the train dataset are colored light brown. Words that were significant on train- *and* holdout datasets are colored dark brown.

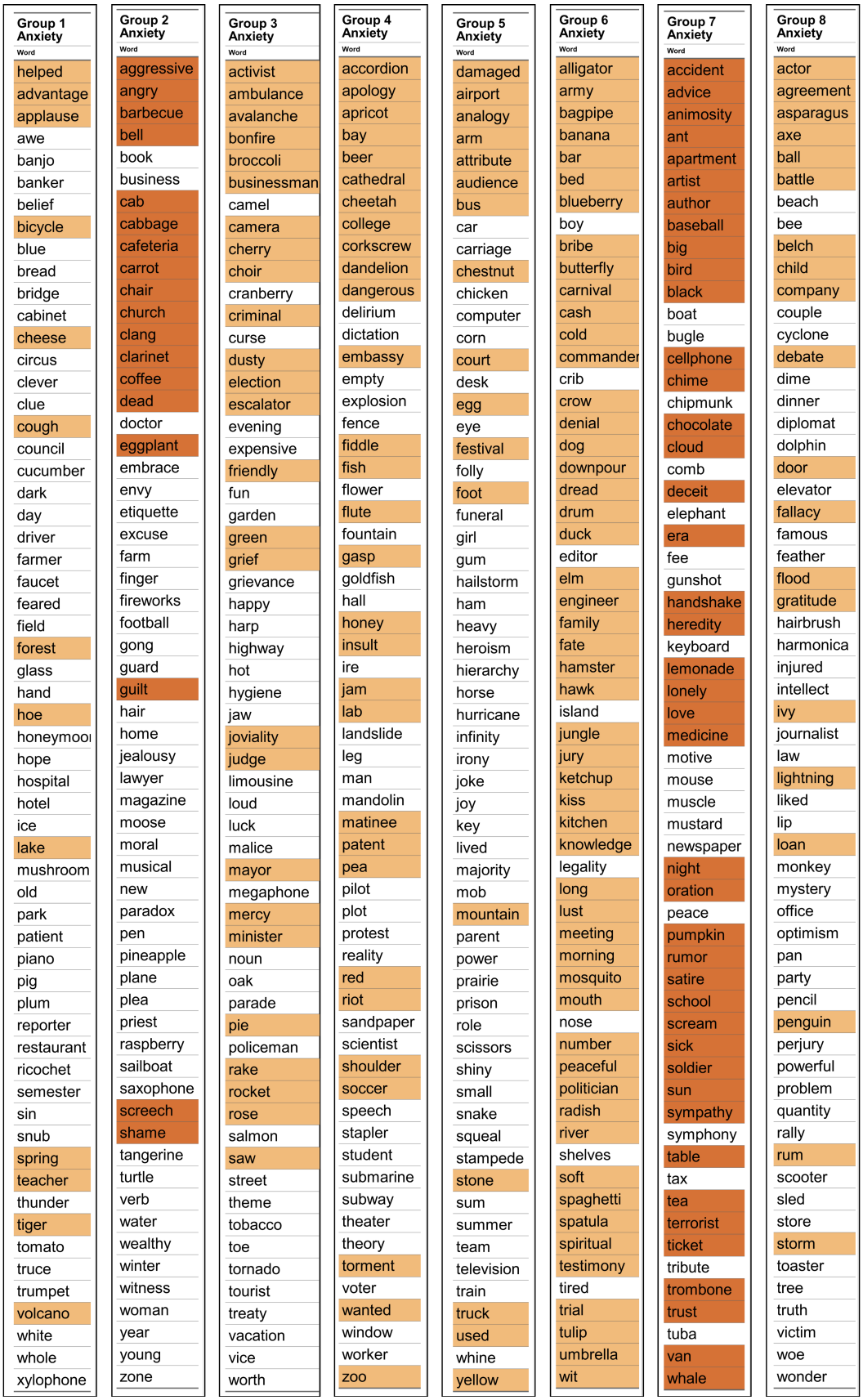


Figure A2: Full list of words for each group. Words that were included in **anxiety**-classifiers that were significant in permutation test 1 on the train dataset are colored light brown. Words that were significant on train- *and* holdout datasets are colored dark brown.

B

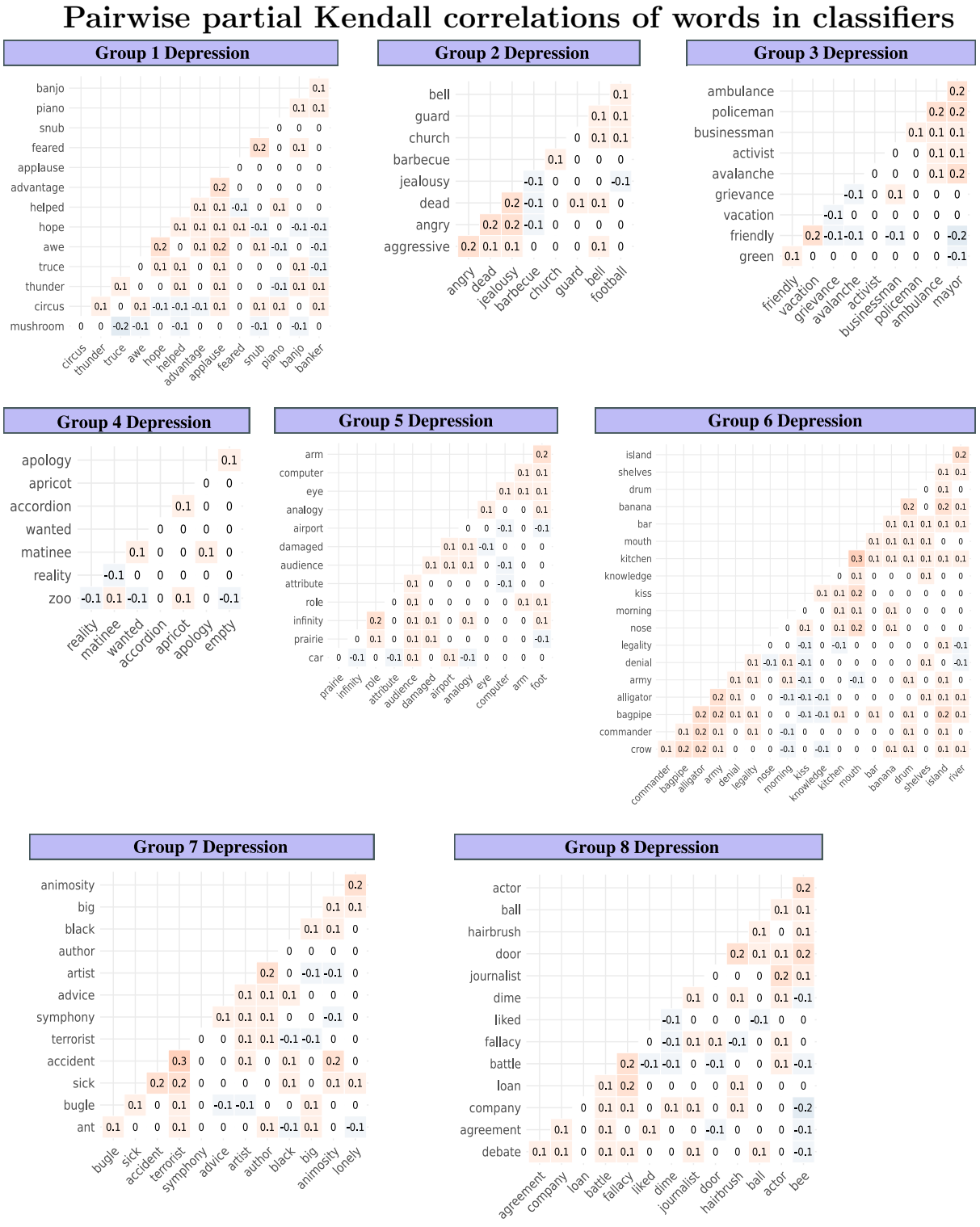


Figure B1: Pairwise partial Kendall correlations of words in **depression**-classifiers.

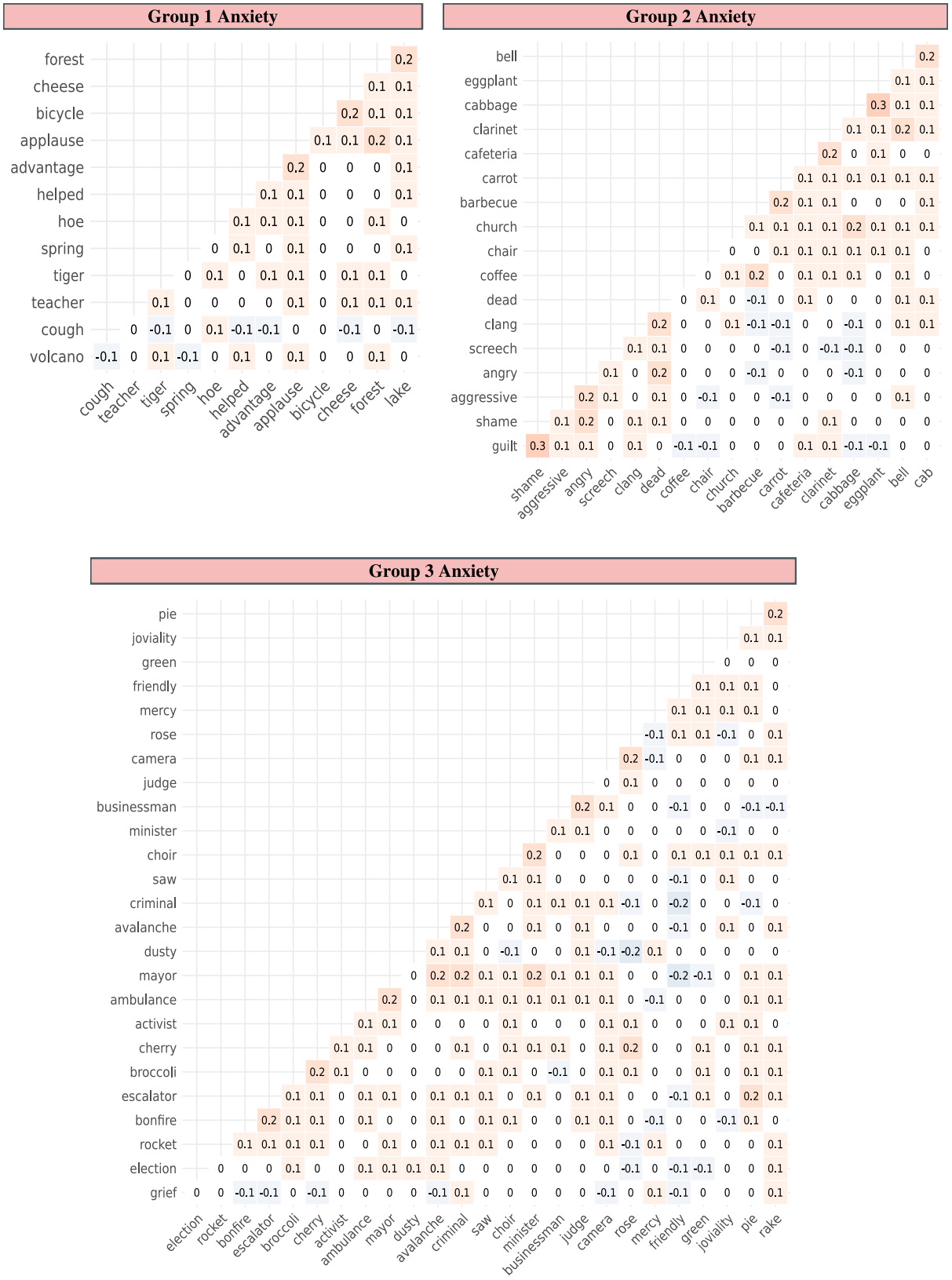


Figure B2: Pairwise partial Kendall correlations of words in **anxiety**-classifiers for groups 1, 2, and 3.

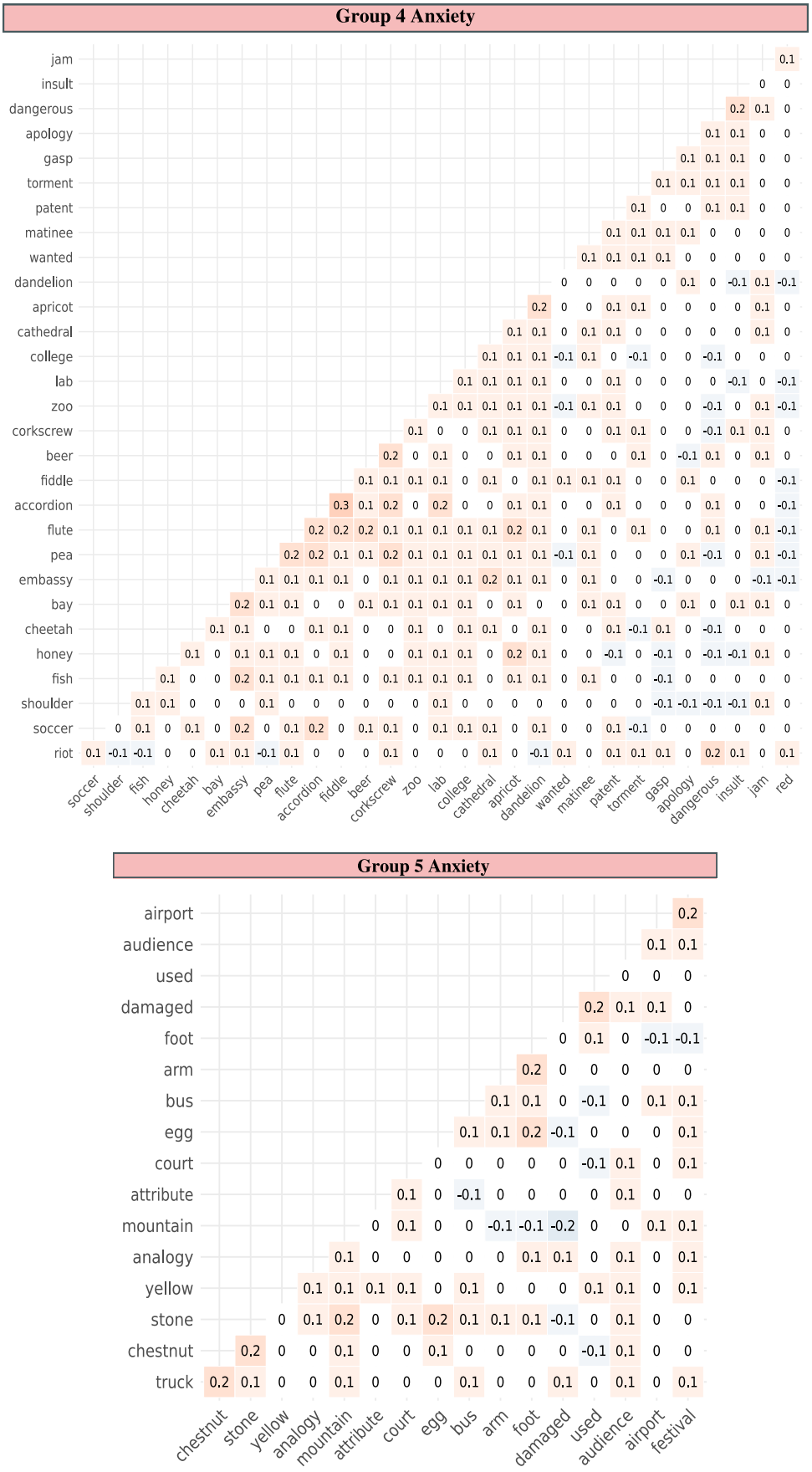


Figure B3: Pairwise partial Kendall correlations of words in **anxiety**-classifiers for groups 4 and 5.

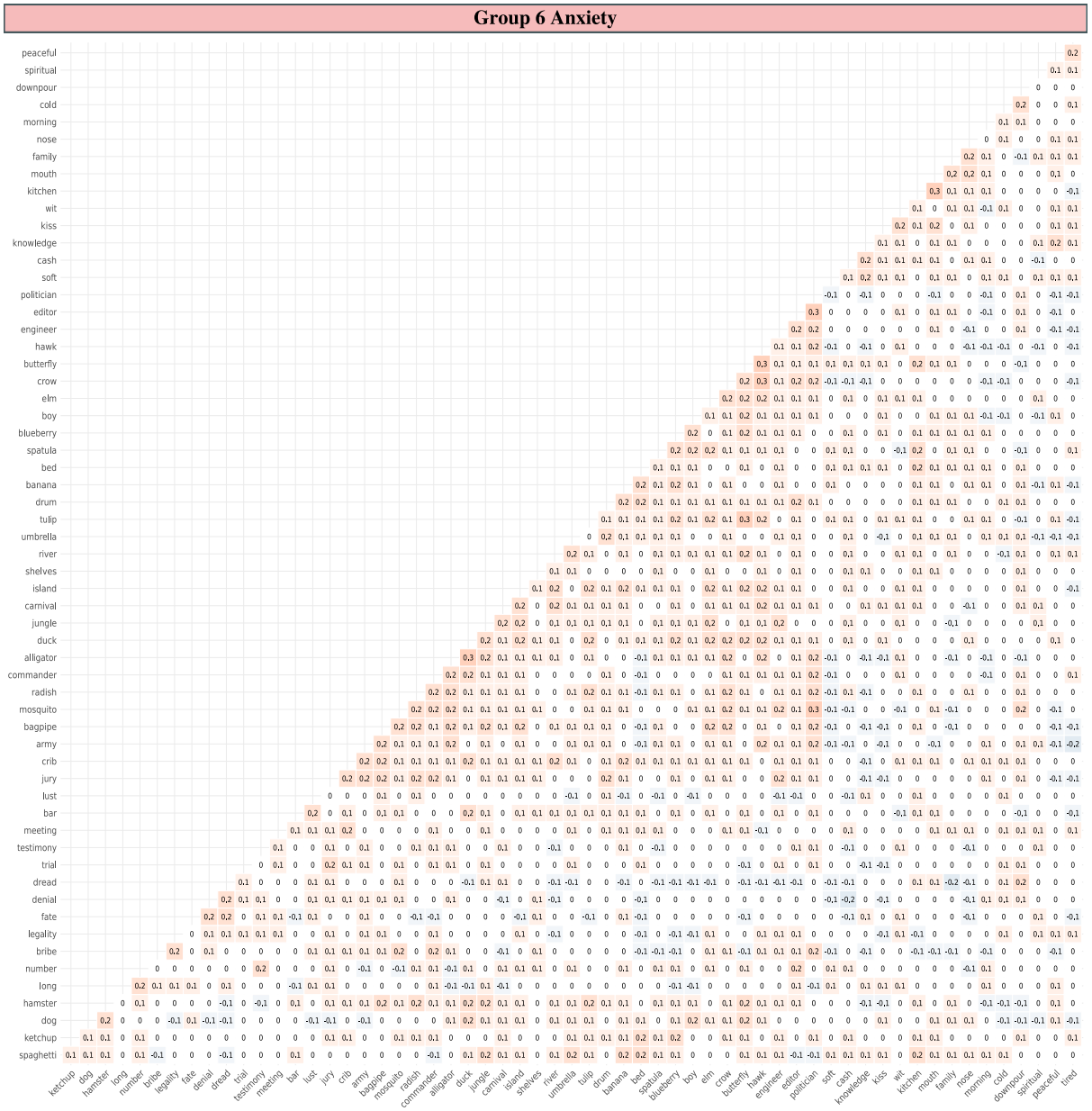


Figure B4: Pairwise partial Kendall correlations of words in the **anxiety**-classifier of group 6.

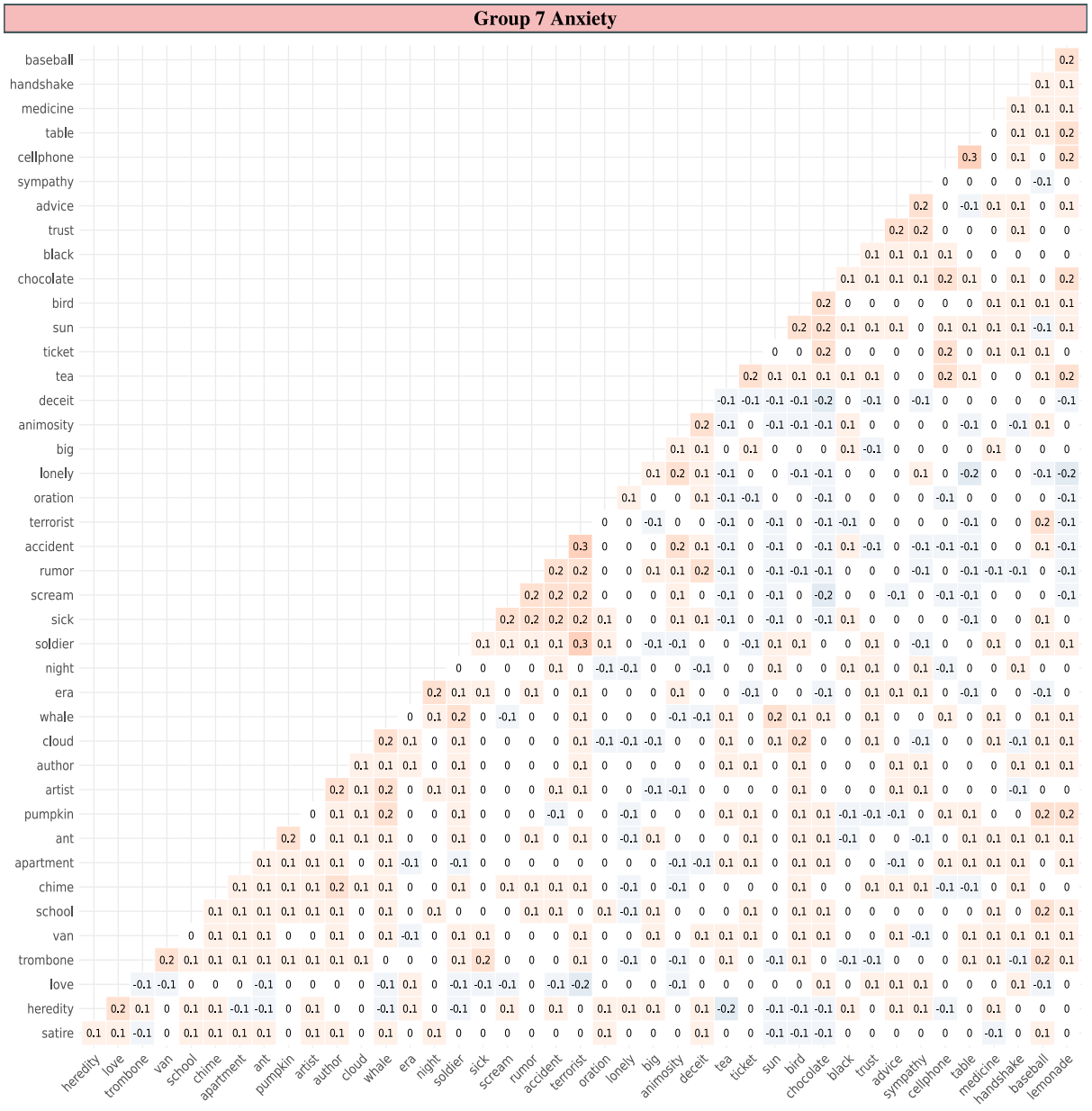


Figure B5: Pairwise partial Kendall correlations of words in the **anxiety**-classifier of group 7.

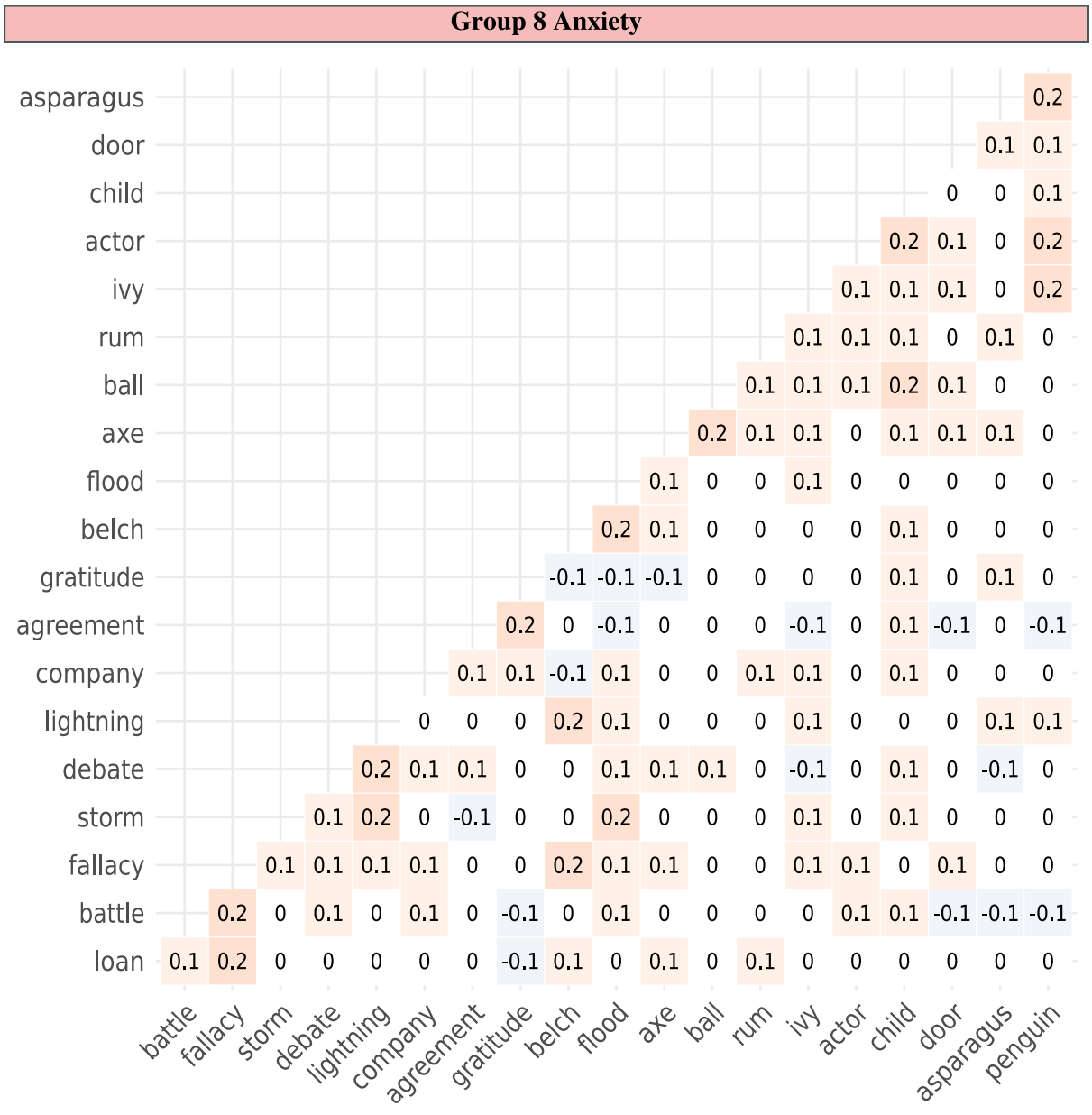


Figure B6: Pairwise partial Kendall correlations of words in the **anxiety**-classifier of group 8.

C

Variance inflation factor scores of words in classifiers

Group 1 Depression

	variable	VIF
0	helped	1.087236
1	advantage	1.054355
2	applause	1.086500
3	awe	1.058793
4	banjo	1.026398
5	banker	1.033004
6	belief	1.038009
7	bridge	1.047929
8	cabinet	1.056613
9	cheese	1.068737
10	dark	1.039473
11	plum	1.080452
12	reporter	1.037964
13	intercept	1.576905

Group 2 Depression

	variable	VIF
0	aggressive	1.082057
1	angry	1.118247
2	barbecue	1.082901
3	bell	1.092932
4	business	1.053042
5	cabbage	1.091395
6	cafeteria	1.161415
7	clarinet	1.145643
8	coffee	1.081697
9	embrace	1.040207
10	farm	1.086425
11	home	1.089575
12	turtle	1.081134
13	verb	1.047323
14	water	1.133307
15	winter	1.092090
16	woman	1.120363
17	intercept	1.999544

Group 3 Depression

	variable	VIF
0	activist	1.030655
1	ambulance	1.128405
2	avalanche	1.098485
3	bonfire	1.048489
4	broccoli	1.047412
5	businessman	1.055186
6	camel	1.137654
7	criminal	1.125544
8	dusty	1.042649
9	grief	1.049387
10	highway	1.065606
11	hygiene	1.037526
12	luck	1.052394
13	minister	1.111516
14	policeman	1.124830
15	intercept	2.108231

Group 4 Depression

	variable	VIF
0	accordion	1.007362
1	apology	1.031517
2	beer	1.018338
3	dictation	1.021040
4	explosion	1.009742
5	theater	1.012701
6	intercept	1.452702

Group 5 Depression

	variable	VIF
0	damaged	1.047159
1	airport	1.015776
2	analogy	1.027537
3	arm	1.046488
4	attribute	1.021477
5	car	1.047971
6	egg	1.040185
7	funeral	1.049538
8	hierarchy	1.031781
9	horse	1.052587
10	role	1.033922
11	intercept	1.444834

Group 6 Depression

	variable	VIF
0	alligator	1.119168
1	army	1.118170
2	bagpipe	1.111506
3	banana	1.038962
4	bar	1.029822
5	carnival	1.084584
6	cold	1.029977
7	family	1.068365
8	hawk	1.112042
9	knowledge	1.049474
10	nose	1.056279
11	river	1.106179
12	testimony	1.039388
13	trial	1.055887
14	umbrella	1.080196
15	intercept	1.972110

Group 7 Depression

	variable	VIF
0	accident	1.108748
1	advice	1.040957
2	animosity	1.131899
3	ant	1.135248
4	artist	1.049566
5	baseball	1.082780
6	lonely	1.062400
7	love	1.064182
8	mouse	1.183131
9	sick	1.088583
10	soldier	1.107133
11	tax	1.035202
12	intercept	1.926874

Group 8 Depression

	variable	VIF
0	actor	1.032136
1	agreement	1.033927
2	ball	1.046551
3	flood	1.024352
4	harmonica	1.032089
5	ivy	1.040092
6	liked	1.022955
7	toaster	1.032187
8	intercept	1.370407

Figure C1: Variance inflation factor scores of words in **depression**-classifiers.

Group 1 Anxiety

	variable	VIF
0	helped	1.093398
1	advantage	1.079414
2	applause	1.102377
3	bicycle	1.056499
4	bread	1.058520
5	circus	1.076451
6	cough	1.083215
7	council	1.076768
8	driver	1.027810
9	hospital	1.064157
10	hotel	1.082365
11	ice	1.055622
12	reporter	1.068709
13	semester	1.093157
14	thunder	1.098265
15	truce	1.052948
16	volcano	1.067763
17	intercept	1.642016

Group 2 Anxiety

	variable	VIF
0	aggressive	1.115208
1	angry	1.125606
2	barbecue	1.166074
3	bell	1.154698
4	book	1.153367
5	business	1.042901
6	cab	1.092580
7	cabbage	1.155516
8	carrot	1.115576
9	chair	1.051796
10	church	1.221080
11	clarinet	1.102195
12	coffee	1.123873
13	dead	1.118247
14	embrace	1.067583
15	gong	1.129980
16	guard	1.118509
17	lawyer	1.053601
18	plea	1.093327
19	sailboat	1.138862
20	winter	1.119764
21	zone	1.086280
22	intercept	2.255336

Group 3 Anxiety

	variable	VIF
0	activist	1.034695
1	ambulance	1.078839
2	avalanche	1.023178
3	camel	1.084974
4	camera	1.043216
5	curse	1.026889
6	escalator	1.081904
7	jaw	1.021457
8	loud	1.020336
9	intercept	1.894930

Group 4 Anxiety

	variable	VIF
0	accordion	1.072970
1	apology	1.029221
2	apricot	1.109506
3	bay	1.052960
4	beer	1.095046
5	cathedral	1.076561
6	cheetah	1.060518
7	college	1.053274
8	corkscrew	1.161221
9	dandelion	1.064272
10	dangerous	1.086520
11	fence	1.081398
12	honey	1.071114
13	patent	1.078720
14	intercept	1.403914

Group 5 Anxiety

	variable	VIF
0	damaged	1.039827
1	airport	1.040786
2	analogy	1.049775
3	arm	1.031895
4	attribute	1.037226
5	audience	1.062137
6	bus	1.085406
7	car	1.063256
8	court	1.014461
9	egg	1.042489
10	funeral	1.051706
11	lived	1.078740
12	intercept	1.550425

Group 6 Anxiety

	variable	VIF
0	alligator	1.228904
1	army	1.198765
2	bagpipe	1.272304
3	banana	1.206853
4	bar	1.155817
5	bed	1.199035
6	blueberry	1.230538
7	boy	1.110322
8	bribe	1.197308
9	butterfly	1.272021
10	carnival	1.159298
11	cash	1.129508
12	cold	1.163342
13	crib	1.273075
14	crow	1.280834
15	downpour	1.179802
16	elm	1.195828
17	fate	1.136963
18	island	1.314677
19	jungle	1.205921
20	jury	1.181698
21	kiss	1.166933
22	kitchen	1.260978
23	legality	1.126111
24	lust	1.173597
25	mosquito	1.297777
26	mouth	1.229940
27	nose	1.147561
28	politician	1.269483
29	radish	1.225308
30	river	1.206963
31	spaghetti	1.217868
32	umbrella	1.176985
33	wit	1.170892
34	intercept	2.555427

Group 7 Anxiety

	variable	VIF
0	accident	1.067917
1	advice	1.028973
2	animosity	1.045214
3	ant	1.035773
4	apartment	1.049147
5	boat	1.056604
6	bugle	1.037217
7	cellphone	1.035662
8	heredity	1.027710
9	night	1.017347
10	intercept	1.676967

Group 8 Anxiety

	variable	VIF
0	actor	1.082305
1	agreement	1.062968
2	asparagus	1.050045
3	axe	1.097052
4	ball	1.131494
5	battle	1.067809
6	beach	1.072503
7	child	1.132376
8	company	1.077670
9	debate	1.093450
10	dime	1.040627
11	dinner	1.112109
12	hairbrush	1.073061
13	ivy	1.082331
14	law	1.082649
15	liked	1.062061
16	loan	1.057169
17	perjury	1.098988
18	storm	1.041708
19	intercept	1.533174

Figure C2: Variance inflation factor scores of words in **anxiety**-classifiers.