Supplementary Information for: What made you do this? Understanding black-box decisions with sufficient input subsets

Contents

S1 Detailed Description of Alternative Methods 3

S2 Details of the Transcription Factor Binding Analysis 3
  S2.1 Dataset and Model ........................................ 3
  S2.2 Rationale length comparison between SIS and other methods ........................................ 4
  S2.3 Evaluation of the quality of TF Rationales ........................................ 5

S3 Details of the MNIST Analysis 6
  S3.1 Dataset and Model ........................................ 6
  S3.2 Local Minima in Backward Selection ........................................ 6
  S3.3 Energy Distance Between Image SIS ........................................ 7
  S3.4 SIS Clustering and Adversarial Analysis ........................................ 7
  S3.5 Understanding Differences Between MNIST Classifiers ........................................ 10

S4 Details of the Beer Reviews Sentiment Analysis 12
  S4.1 Beer Reviews Data Description ........................................ 12
  S4.2 Model Architecture and Training ........................................ 12
  S4.3 Imputation Strategies: Mean vs. Hot-deck ........................................ 13
  S4.4 Feature Importance Scores ........................................ 13
  S4.5 Additional Results for Aroma aspect ........................................ 13
  S4.6 Understanding Differences Between Sentiment Predictors ........................................ 14
  S4.7 Results for Appearance and Palate aspects ........................................ 19

List of Figures

1 Beer review with SIS for each aspect ........................................ 5
2 Beer review with 3 SIS subsets ........................................ 5
3 Prediction on rationales only vs. rationale length (aroma prediction) ........................................ 5
4 QHS vs. SIS-annotation similarity ........................................ 5
5 Example DNA sequences ........................................ 6
6 KL divergence for TF motifs and rationales ........................................ 6
7 SIS clusters for digit 4 ........................................ 7
8 SIS-collections for adversarial and misclassified MNIST digits ........................................ 7
9 Joint clustering MNIST digit 4 SIS from CNN & MLP ........................................ 7
10 Predictions on the SIS from alternative models on beer reviews and MNIST digits ........................................ 8
S1 AUC performance of CNN models on TF binding prediction ........................................ 4
S2 TF dataset sufficiency thresholds ........................................ 4
S3 Length of rationales in TF binding prediction ........................................ 5
S4 TF task rationale only prediction comparison ........................................ 5
S5 Local Minimum in Backward Selection ........................................ 6
S6 Number of examples per MNIST digit ........................................ 7
S7 Number of SIS per MNIST digit ........................................ 7
Understanding black-box decisions with sufficient input subsets

S8 SIS examples on MNIST (CNN) .......................................................... 8
S9 SIS clusters identified on MNIST (CNN) ............................................. 9
S10 SIS clusters identified on MNIST (MLP) ........................................... 10
S11 SIS examples on MNIST (MLP) .......................................................... 11
S12 Mean vs. hot-deck imputation for aroma prediction ............................ 13
S13 Feature importance comparison for aroma prediction .......................... 14
S14 Length of rationales for aroma prediction ......................................... 14
S15 Prediction of aroma sentiment on the annotation set ............................ 15
S16 Number of SIS per for aroma beer reviews ........................................ 15
S17 Prediction as function of remaining text for aroma prediction ............... 15
S18 Alignment of human rationales for beer reviews with predictive model .... 15
S19 Mean vs. hot-deck imputation in appearance prediction ........................ 19
S20 Prediction of appearance sentiment on the annotation set ...................... 19
S21 Number of SIS per for appearance beer reviews .................................. 19
S22 Length of rationales for appearance prediction .................................... 20
S23 Feature importance for appearance prediction .................................... 20
S24 QHS vs. fraction of SIS in human rationale for appearance prediction .... 20
S25 Prediction on only rationale vs. rationale length .................................. 20
S26 Prediction as function of remaining text for appearance aspect .............. 21
S27 Mean vs. hot-deck imputation for palate prediction ............................. 22
S28 Prediction of palate sentiment on the annotation set ............................ 22
S29 Number of SIS per for palate beer reviews ........................................ 22
S30 Length of rationales for palate prediction .......................................... 22
S31 Feature importance for palate prediction ......................................... 22
S32 QHS vs. fraction of SIS in human rationale for palate prediction .......... 23
S33 Prediction on only rationale vs. rationale length for palate prediction ...... 23
S34 Prediction as function of remaining text for palate aspect .................... 23

List of Tables

1 Three SIS clusters from beer reviews ................................................... 8
2 Joint clustering of SIS from LSTM and CNN on beer reviews, aroma aspect 8
S1 Summary and performance statistics for beer reviews data ...................... 12
S2 Statistics for rationale length and feature importance in aroma prediction 14
S3 SIS clusters for positive aroma prediction .......................................... 16
S4 SIS clusters for negatives aroma prediction ........................................ 16
S5 Joint clustering of SIS from LSTM and CNN on beer reviews, positive aroma aspect 17
S6 Joint clustering of SIS from LSTM and CNN on beer reviews, negative aroma aspect 18
S7 Statistics for rationale length and feature importance in appearance prediction 20
S8 SIS clusters for positive appearance prediction ...................................... 21
S9 SIS clusters for negative appearance prediction .................................... 21
S10 Statistics for rationale length and feature importance in palate prediction 21
S11 SIS clusters for positive palate prediction .......................................... 23
S12 SIS clusters for negative palate prediction ........................................... 24
S1 Detailed Description of Alternative Methods

In Section 3, we describe a number of alternative methods for identifying rationales for comparison with our method. We use methods based on integrated gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016), and feature perturbation. Note that integrated gradients is an attribution method which assigns a numerical score to each input feature. LIME likewise assigns a weight to each feature using a local linear regression model for \( f \) around \( \mathbf{x} \). In the perturbative approach, we compute the change in prediction when each feature is individually masked, as in Equation 1 (of Section S4.4). Each of these feature orderings \( R \) is used to construct a rationale using the FindSIS procedure (Section 3) for the “Suff. IG,” “Suff. LIME,” and “Suff. Perturb.” (sufficiency constrained) methods.

Note that our text classification architecture (described in Section S4.2) encodes discrete words as 100-dimensional continuous word embeddings. The integrated gradients method returns attribution scores for each coordinate of each word embedding. For each word embedding \( x_i \in \mathbf{x} \) (where each \( x_i \in \mathbb{R}^{100} \)), we summarize the attributions along the corresponding embedding into a single score \( y_i \) using the \( L_1 \) norm: \( y_i = \sum_d |x_{id}| \) and compute the ordering \( R \) by sorting the \( y_i \) values.

We use an implementation of integrated gradients for Keras-based models from https://github.com/hiranum/IntegratedGradients. In the case of the beer review dataset (Section 4.1), we use the mean embedding vector as a baseline for computing integrated gradients. In the case of TF binding (Section 4.2), we use the \([0.25, 0.25, 0.25, 0.25]\) uniform mean vector as the baseline reference value. As suggested in Sundararajan et al. (2017), we verified that the prediction at the baseline and the integrated gradients sum to approximately the prediction of the input.

For LIME and our beer reviews dataset, we use the approach described in Ribeiro et al. (2016) for textual data, where individual words are removed entirely from the input sequence. In our TF binding dataset, LIME replaces bases with the unknown \( N \) base (represented as the uniform-distribution \([0.25, 0.25, 0.25, 0.25]\)). We use the implementation of LIME at: https://github.com/marcotcr/lime. The LimeTextExplainer module is used with default parameters, except we set the maximal number of features used in the regression to be the full input length so we can order all input features.

Additionally, we explore methods in which we use the same ordering \( R \) by these alternative methods but select the same number of input features in the rationale to be the median SIS length in the SIS-collection computed by our method on each example: the “IG,” “LIME,” and “Perturb.” (length constrained) methods. In the TF binding models, we use a baseline of zero vectors such that the integrated gradients result along the encoded sequence is also one-hot. We compute the feature ordering based on the absolute value of the non-zero integrated gradient attributions.

In TF binding data (Section 4.2), we add an additional method, “Top IG,” in which we compute integrated gradients using an all-zeros baseline and order features by attribution magnitude (as in the length constrained IG method). But, we select elements for the rationale by finding the minimum number of elements necessary such that the sum of integrated gradients of those features equals \( \tau - f(\mathbf{0}) \), where \( \mathbf{0} \) is the all-zeros baseline for integrated gradients. Note that for the length constrained and Top IG methods, there is no guarantee of sufficiency \( f(\mathbf{x}_S) \geq \tau \) for any input subset \( S \).

S2 Details of the Transcription Factor Binding Analysis

S2.1 Dataset and Model

We use the motif occupancy datasets\(^1\) from Zeng et al. (2016), where each dataset originates from a ChIP-seq experiment from the ENCODE project (Consortium et al., 2012). Each of the 422 datasets studies a particular transcription factor, containing between 600 and 700,000 (median 50,000) 101 base-pair DNA sequences (inputs) each associated with a binary label based on whether the sequence is bound by the TF or not. Each dataset also contains a test set ranging between 150 and 170,000 sequences (median 12,000). Here, the positive and negative classes in each dataset are balanced, and we filter out all sequences containing the unknown base (\( N \)).

\(^1\)available at http://cnn.csail.mit.edu
Understanding black-box decisions with sufficient input subsets

Figure S1: Median area under the receiver operating curve (AUC) for all 422 transcription factor binding motif occupancy datasets. The validation set is held-out at training but used to choose model parameters; the test set is not seen until after training.

Figure S2: Thresholds $\tau$ used for identifying sufficient input subsets in TF binding datasets. In each dataset, the threshold is defined as the 90th percentile of the predictive test distribution.

The nucleotide occurring at base position (A, C, G, T) is encoded as a one-hot representation which is fed into the CNN. Zeng et al. (2016) showed that convolutional neural network architectures outperform other models for this TF binding prediction task.

For each of the 422 prediction tasks, we employ the best-performing “1layer_128motif” architecture from Zeng et al. (2016), defined as follows:

1. **Input:** (101 x 4) sequence encoding
2. **Convolutional Layer 1:** Applies 128 kernels of window size 24, with ReLU activation
3. **Global Max Pooling Layer 1:** Performs global max pooling
4. **Dense Layer 1:** 32 neurons, with ReLU activation and dropout probability 0.5
5. **Dense Layer 2:** 1 neuron (output probability), with sigmoid activation

We hold out 1/8 of each train set for validation and minimize binary cross-entropy using the Adadelta optimizer (Zeiler, 2012) with default parameter settings in Keras (Chollet et al., 2015). We train each model on each of the 422 datasets for 10 epochs (using batch size 128) with early-stopping based on validation loss. Figure S1 shows the area under the receiver operating curve (AUC) over the 422 datasets, and we note that the performance of our models closely resembles that in Zeng et al. (2016).

**S2.2 Rationale length comparison between SIS and other methods**

For each dataset, we define the sufficiency threshold $\tau$ as the 90th percentile of the predictive distribution on all test sequences. The distribution of thresholds is shown in Figure S2. We compute the complete set of sufficient input subsets for each corresponding test sequence. Since A,C,G,T nucleotides all occur with similar frequency in this data, our SIS analysis simply masks each base using a uniform embedding ([0.25, 0.25, 0.25, 0.25]). This is also the standard strategy to represent unknown “N” nucleotides in DNA sequences that typically arise from issues in read quality. We generally find that there is only a single SIS per example for the sequences in these datasets.

On each dataset, we compute the median rationale length (as number of bases in the rationale). The distribution of median rationale length over all datasets by various methods is shown in Figure S3. Note that for the IG, LIME, and Perturb. methods, rationale length was constrained to the length of the rationales produced by our method. For the Top IG method, neither sufficiency or length constraints are enforced. We see that when the sufficiency constraint is enforced in alternative methods (Suff. IG), the rationales are significantly longer than
Figure S3: Length (number of bases) of rationales identified by various methods. Note that the sufficiency constraint \( f(x_S) \geq \tau \) is only enforced for SIS and Suff. IG. The lengths of IG, LIME, and Perturb. rationales are constrained to the length of SIS rationales.

those identified by SIS. Moreover, as shown in Figure S4, when the sufficiency constraint is not enforced (or the rationale lengths are constrained to the length of SIS rationales) in alternative methods, the rationales have significantly less predictive power, often not satisfying \( f(x_S) \geq \tau \).

S2.3 Evaluation of the quality of TF Rationales

Each rationale is padded with “\( \mathbb{N} \)” (unknown) bases to the length of a full input sequence (101 bases) and optimally aligned with the known motif\(^2\) according to the likelihood criterion. The aligned motif is then also padded to the same length, and we compute the divergence between between the rationale \( R \) and known motif \( M \) as:

\[
\text{Div}(R, M) = \sum_i D_{\text{KL}}(R_i \| M_i)
\]

where \( D_{\text{KL}}(R_i \| M_i) = \sum_j R_i(j) \log \frac{R_i(j)}{M_i(j)} \) is the Kullback-Leibler divergence from \( M_i \) to \( R_i \), and \( M_i \) and \( R_i \) are distributions over bases (A, C, G, T) at position \( i \). Note that as \( R \) and \( M \) become more dissimilar, \( \text{Div}(R, M) \) increases. We ensure \( M_{ij} > 0 \ \forall \ i, j \) so \( D_{\text{KL}} \) is always finite.

\(^2\)A JASPAR motif is a \( n \times 4 \) right stochastic matrix \( M \). The columns represent the ACGT DNA bases and the rows a DNA sequence. It represents the marginal probability of the base \( j \) at position \( i \) being present with probability \( M_{ij} \). The unknown base “\( \mathbb{N} \)” receives uniform 1/4 probability for each of ACGT.
S3 Details of the MNIST Analysis

S3.1 Dataset and Model

The MNIST database of handwritten digits contains 60k training images and 10k test images (LeCun et al., 1998). All images are 28x28 grayscale, and we normalize them such that all pixel values are between 0 and 1. We use the convolutional architecture provided in the Keras MNIST CNN example. The architecture is as follows:

1. **Input**: (28 x 28 x 1) image, all values $\in [0, 1]$
2. **Convolutional Layer 1**: Applies 32 3x3 filters with ReLU activation
3. **Convolutional Layer 2**: Applies 64 3x3 filters, with ReLU activation
4. **Pooling Layer 1**: Performs max pooling with a 2x2 filter and dropout probability 0.25
5. **Dense Layer 1**: 128 neurons, with ReLU activation and dropout probability 0.5
6. **Dense Layer 2**: 10 neurons (one per digit class), with softmax activation

The Adadelta optimizer (Zeiler, 2012) is used to minimize cross-entropy loss on the training set. The final model achieves 99.7% accuracy on the train set and 99.1% accuracy on the held-out test set.

S3.2 Local Minima in Backward Selection

Figure S5: (a) Prediction on remaining image as pixels are masked during backward selection, when our CNN classifier is fed the MNIST digit in (b). The dashed line depicts the threshold $\tau = 0.7$. (b) Original image (class 9). (c) SIS if backward selection were to terminate the first time prediction on remaining image drops below 0.7, corresponding to point C in (a) (CNN predicts class 9 with probability 0.700 on this SIS). (d) Actual SIS produced by our FindSIS algorithm, corresponding to point D in (a) (CNN predicts class 9 with probability 0.704 on this SIS).

Figure S5 demonstrates an example MNIST digit for which there exists a local minimum in the backward selection phase of our algorithm to identify the initial SIS. Note that if we were to terminate the backward selection as soon as predictions drop below the decision threshold, the resulting SIS would be overly large, violating our minimality criterion. It is also evident from Figure S5 that the smaller-cardinality SIS in (d), found after the initial local optimum in (c), presents a more interpretable input pattern that enables better understanding of the core motifs influencing our classifier’s decisions. To avoid suboptimal results, it is important to run a complete backward selection sweep until the entire input is masked before building the SIS upward, as done in our SIScollection procedure.

---

3http://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py
S3.3  Energy Distance Between Image SIS

To cluster SIS from the image data, we compute the pairwise distance between two SIS subsets $S_1$ and $S_2$ as the energy distance (Rizzo and Székely, 2016) between two distributions over the image pixel coordinates that comprise the SIS, $X_1$ and $X_2 \in \mathbb{R}^2$:

$$D(X_1, X_2) = 2 \cdot \mathbb{E} \left[ \|X_1 - X_2\| - \mathbb{E} \left[ \|X_1 - X'_1\| - \mathbb{E} \left[ \|X_2 - X'_2\| \right] \right] \right] \geq 0$$

Here, $X_i$ is uniformly distributed over the pixels that are selected as part of the SIS subset $S_i$, $X'_i$ is an i.i.d. copy of $X_i$, and $\| \cdot \|$ represents the Euclidean norm. Unlike a Euclidean distance between images, our usage of the energy distance takes into account distances between the similar pixel coordinates that comprise each SIS. The energy distance offers a more efficiently computable integral probability metric than the optimal transport distance, which has been widely adopted as an appropriate measure of distance between images.

S3.4  SIS Clustering and Adversarial Analysis

We set the threshold $\tau = 0.7$ for SIS to ensure that the model is confident in its class prediction (probability of the predicted class is $\geq 0.7$). Almost all test examples initially have $f(x) \geq \tau$ for the top class (Figure S6). We identify all test examples that satisfy this condition and use SIS to identify all sufficient input subsets. The number of sufficient input subsets per digit is shown in Figure S7.

We apply our SIScollection algorithm to identify sufficient input subsets on MNIST test digits (Section 4.3). Examples of the complete SIS-collection corresponding to randomly chosen digits are shown in Figure S8. We also cluster all the sufficient input subsets identified for each class (Section 4.4), depicting the results in Figure S9.

In Figure 8, we show an MNIST image of the digit 9, adversarially perturbed to 4, and the sufficient subsets corresponding to the adversarial prediction. Although a visual inspection of the perturbed image does not really reveal exactly how it has been manipulated, it becomes immediately clear from the SIS-collection for the adversarial image. These sets shows that the perturbation modifies pixels in such a way that input patterns similar to the typical SIS-collection for a 4 (Figure 7) become embedded in the image. The adversarial manipulation was done using the Carlini-Wagner $L_2$ (CW2) attack$^4$ (Carlini and Wagner, 2017b) with a confidence parameter of 10. The CW2 attack tries to find the minimal change to the image, with respect to the $L_2$ norm, that will lead the image to be misclassified. Carlini and Wagner (2017a) demonstrate it to be one of the strongest extant adversarial attacks.

$^4$Implemented in the cleverhans library of Papernot et al. (2017)
Figure S8: Visualization of SIS-collections identified from MNIST digits that are confidently classified by the CNN. For each class, six examples were chosen randomly. For each example, we show the original image (left) and the complete set of sufficient input subsets identified for that example (remaining images in each row). Each individual SIS satisfies $f(x_S) \geq \tau$ for that class.
Figure S9: Clustering all the SIS found for each digit under the CNN model (see Section 4.4). Each row contains images drawn from one cluster. The bottom row (“Misc”) contains a sample of miscellaneous SIS not assigned to any cluster by DBSCAN.
S3.5 Understanding Differences Between MNIST Classifiers

We use SIS and our clustering procedure to understand and visualize differences in features learned by two different models trained on the same MNIST digit classification task. In addition to the previously-described CNN model (see Section S3.1), we also trained a simple multilayer perceptron (MLP) on the same task. The MLP architecture is as follows:

1. **Input**: 784-dimensional (flattened) image, all values $\in [0,1]$
2. **Dense Layer 1**: 250 neurons, ReLU activation, and dropout probability 0.2
3. **Dense Layer 2**: 250 neurons, ReLU activation, and dropout probability 0.2
4. **Dense Layer 3**: 10 neurons (one per digit class), with softmax activation

As with the CNN, Adadelta (Zeiler, 2012) is used to minimize cross-entropy loss on the training set. The final MLP model achieves 99.7% accuracy on the train set and 98.3% accuracy on the test set, which is close to the performance of the CNN (see Section S3.1).

We apply the same procedure as in Section 4.3 to extract the SIS-collection from all applicable test images using the MLP. To understand differences between the feature patterns that each model has learned to associate with predicting each digit, we combine all SIS (from both models for a particular class) and run our clustering procedure (see Section 4.4 and Figure 9). In the resulting clustering, we list what percentage of the SIS in each cluster stem from the CNN vs. the MLP. Most clusters contain examples purely from a single model, indicating the two models have learned to associate different feature patterns with the target class (Figure 9), which was chosen to be the digit 4 in this case.

For further comparison, we include clustering results for the SIS extracted from the MLP as evidence for digits 4 and 7 (Figure S10). Additionally, Figure S11 shows all of the SIS extracted from example digits from these classes applying our procedure on the MLP.

![Figure S10: Clustering all the SIS identified by our method on digits 4 and 7 under the MLP model (see Section 4.4). Each row contains images drawn from one cluster. The bottom row ("Misc") contains a sample of miscellaneous SIS not assigned to any cluster by DBSCAN. Compare to the SIS-clustering from our CNN model (Figure S9).](image)
Figure S11: Visualization of SIS-collections identified for MNIST digits 4 and 7 under the MLP model. For each class, six examples were chosen randomly. For each example, we show the original image (left) and the complete set of sufficient input subsets identified for that example (remaining images in each row). Note that each individual SIS satisfies $f(x_S) \geq \tau$ for that class. Compare to the SIS extracted from our CNN (Figure S8).
S4 Details of the Beer Reviews Sentiment Analysis

S4.1 Beer Reviews Data Description

Following Lei et al. (2016), we use a preprocessed version of the BeerAdvocate dataset which contains decorrelated numerical ratings toward three aspects: aroma, appearance, and palate (each normalized to [0, 1]). Dataset statistics can be found in Table S1. Reviews were tokenized by converting to lowercase and filtering punctuation, and we used a vocabulary containing the top 10,000 most common words. McAuley et al. (2012) also provide a subset of human-annotated reviews, in which humans manually selected full sentences in each review that describe the relevant aspects. This annotated set was never seen during training and used solely as part of our evaluation.

S4.2 Model Architecture and Training

Long short-term memory (LSTM) networks are commonly employed for natural language tasks such as sentiment analysis (Wang et al., 2016; Radford et al., 2017). We use a recurrent neural network (RNN) architecture with two stacked LSTMs as follows:

1. **Input/Embeddings Layer**: Sequence with 500 timesteps, the word at each timestep is represented by a (learned) 100-dimensional embedding

2. **LSTM Layer 1**: 200-unit recurrent layer with LSTM (forward direction only)

3. **LSTM Layer 2**: 200-unit recurrent layer with LSTM (forward direction only)

4. **Dense**: 1 neuron (sentiment output), sigmoid activation

With this architecture, we use the Adam optimizer (Kingma and Ba, 2015) to minimize mean squared error (MSE) on the training set. We use a held-out set of 3,000 examples for validation (sampled at random from the pre-defined test set from Lei et al. (2016)). Our test set consists of the remaining 7,000 test examples. Training results are shown in Table S1.

Table S1: Summary and performance statistics (mean squared error (MSE) and Pearson correlation coefficient $\rho$) for beer reviews data and LSTM models.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Fold</th>
<th>Size</th>
<th>MSE</th>
<th>Pearson $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance</td>
<td>Train</td>
<td>80,000</td>
<td>0.016</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>3,000</td>
<td>0.024</td>
<td>0.783</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>7,000</td>
<td>0.023</td>
<td>0.801</td>
</tr>
<tr>
<td></td>
<td>Annotation</td>
<td>994</td>
<td>0.020</td>
<td>0.563</td>
</tr>
<tr>
<td>Aroma</td>
<td>Train</td>
<td>70,000</td>
<td>0.014</td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>3,000</td>
<td>0.024</td>
<td>0.767</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>7,000</td>
<td>0.025</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Annotation</td>
<td>994</td>
<td>0.021</td>
<td>0.598</td>
</tr>
<tr>
<td>Palate</td>
<td>Train</td>
<td>70,000</td>
<td>0.016</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>3,000</td>
<td>0.029</td>
<td>0.680</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>7,000</td>
<td>0.028</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>Annotation</td>
<td>994</td>
<td>0.016</td>
<td>0.592</td>
</tr>
</tbody>
</table>

5https://www.beeradvocate.com/
6http://snap.stanford.edu/data/web-BeerAdvocate.html
S4.3 Imputation Strategies: Mean vs. Hot-deck

In Section 3, we discuss the problem of masking input features. Here, we show that the mean-imputation approach (in which missing inputs are masked with a mean embedding, taken over the entire vocabulary) produces a nearly identical change in prediction to a nondeterministic hot-deck approach (in which missing inputs are replaced by randomly sampling feature-values from the data). Figure S12 shows the change in prediction $f(\mathbf{x}\setminus\{i\}) - f(\mathbf{x})$ by both imputation techniques after drawing a training example $\mathbf{x}$ and word $x_i \in \mathbf{x}$ (both uniformly at random) and replacing $x_i$ with either the mean embedding or a randomly selected word (drawn from the vocabulary, based on counts in the training corpus). This procedure is repeated 10,000 times. Both resulting distributions have mean near zero ($\mu_{\text{mean-embedding}} = -7.0e-4$, $\mu_{\text{hot-deck}} = -7.4e-4$), and the distribution for mean embedding is slightly narrower ($\sigma_{\text{mean-embedding}} = 0.013$, $\sigma_{\text{hot-deck}} = 0.018$). We conclude that mean-imputation is a suitable method for masking information about particular feature values in our SIS analysis.

We also explored other options for masking word information, e.g. replacement with a zero embedding, replacement with the learned $<$PAD$>$ embedding, and simply removing the word entirely from the input sequence, but each of these alternative options led to undesirably larger changes in predicted values as a result of masking, indicating they appear more informative to $f$ than replacement via the feature-mean.

![Change in prediction](image)

Figure S12: Change in prediction $(f(\mathbf{x}\setminus\{i\}) - f(\mathbf{x}))$ after masking a randomly chosen word with mean imputation or hot-deck imputation. 10,000 replacements were sampled from the aroma beer reviews training set.

S4.4 Feature Importance Scores

For each feature $i$ in the input sequence, we quantify its marginal importance by individually perturbing only this feature:

$$\text{Feature Importance}(i) = \text{prediction on original input} - \text{prediction with feature } i \text{ masked}$$

Note that these marginal Feature Importance scores are identical to those of the Perturb. method described in Section S1. The marginal Feature Importance scores are summarized in Table S2 and Figure S13. Compared to the Suff. IG and Suff. LIME methods, our SIScollection technique produces rationales that are much shorter and contain fewer irrelevant (i.e. not marginally important) features (Table S2, Figures S13 and S14). Note that by construction, the rationales of the Suff. Perturb. method contain features with the greatest Feature Importance, since this precisely how the ranking in Suff. Perturb. is defined.

S4.5 Additional Results for Aroma aspect

We apply our method to the set of reviews containing sentence-level annotations. Note that these reviews (and the human annotations) were not seen during training. We choose thresholds $\tau_+ = 0.85$, $\tau_- = 0.45$ for strong positive and strong negative sentiment, respectively, and extract the complete set of sufficient input subsets using our method. Note that in our formulation above, we apply our method to inputs $\mathbf{x}$ where $f(\mathbf{x}) \geq \tau$. For
Table S2: Statistics for rationale length and feature importance in aroma prediction. For rationale length, median and max indicate percentage of input text in the rationale. For marginal perturbed feature importance, we indicate the median importance of features in rationales and features from the other (non-rationale) text. \(p\)-values are computed using a Wilcoxon rank-sum test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rationale Length (% of text)</th>
<th>Marginal Perturbed Feature Importance</th>
<th>(p) (vs. SIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med.</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>3.9%</td>
<td>17.3%</td>
<td>–</td>
</tr>
<tr>
<td>Suff. IG</td>
<td>7.7%</td>
<td>89.7%</td>
<td>5e-26</td>
</tr>
<tr>
<td>Suff. LIME</td>
<td>7.2%</td>
<td>84.0%</td>
<td>4e-23</td>
</tr>
<tr>
<td>Suff. Perturb.</td>
<td>5.1%</td>
<td>18.3%</td>
<td>1e-06</td>
</tr>
</tbody>
</table>

Figure S13: Importance of individual features in the rationales for aroma prediction in beer reviews

Figure S14: Length of rationales for aroma prediction

In the sentiment analysis task, we analogously apply our method for both \(f(x) \geq \tau_+\) and \(-f(x) \geq -\tau_-\), where the model predicts either strong positive or strong negative sentiment, respectively. These thresholds were set empirically such that they were sufficiently apart, based on the distribution of predictions (Figure S15). For most reviews, SISC\textsc{ollection} outputs just one or two SIS sets (Figure S16).

We analyzed the predictor output following the elimination of each feature in the \textsc{BackSelect} procedure (Section 3). Figure S17 shows the LSTM output on the remaining unmasked text \(f(x_{\text{\textsc{U}}}^{(\ell)}))\) at each iteration of \textsc{BackSelect}, for all examples. This figure reveals that only a small number of features are needed by the model in order to make a strong prediction (most features can be removed without changing the prediction). We see that as those final, critical features are removed, there is a rapid, monotonic decrease in output values. Finally, we see that the first features to be removed by \textsc{BackSelect} are those which generally provide negative evidence against the decision.

S4.6 Understanding Differences Between Sentiment Predictors

We demonstrate how our SIS-clustering procedure can be used to understand differences in the types of concepts considered important by different neural network architectures. In addition to the LSTM (see Section S4.2), we trained a convolutional neural network (CNN) on the same sentiment analysis task (on the aroma aspect). The CNN architecture is as follows:

1. \textbf{Input/Embeddings Layer}: Sequence with 500 timesteps, the word at each timestep is represented by a (learned) 100-dimensional embedding
2. \textbf{Convolutional Layer 1}: Applies 128 filters of window size 3 over the sequence, with ReLU activation
3. \textbf{Max Pooling Layer 1}: Max-over-time pooling, followed by flattening, to produce a (128,) representation
4. \textbf{Dense}: 1 neuron (sentiment output), sigmoid activation
S4.6 Understanding Differences Between Sentiment Predictors

Figure S15: Predictive distribution on the annotation set (held-out) using the LSTM model for aroma. Vertical lines indicate decision thresholds ($\tau_+ = 0.85$, $\tau_- = 0.45$) selected for SIScollection.

Figure S16: Number of sufficient input subsets for aroma identified by SIScollection per example.

Figure S17: Prediction history on remaining (unmasked) text at each step of the BackSelect procedure, for examples where aroma sentiment is predicted.

**Good Alignment with Human-selection:** QHA = 0.00296, Full Sequence Prediction = 0.900

> poured from a 24oz bottle into a large appearance this pours a deep bronze amber in color this beer has some of the best head formation and retention that i have ever seen along with lots and lots of sticky lacing smell tons of honey resonant balanced abound in this ale cant get enough of this beer its the best smelling harvest ale i ve ever had the pleasure of enjoying taste huge flavor profile with lots of bitterness and only a little malt sweetness as this beer warms up the bitterness really dominates the flavor i m tasting lot of subtle orange aromas mouthfeel full body beer with alot of carbonation overall i love this beer in my opinion its nevada s best beer overall the price is i bought my for 3.99 so much flavor and biting bitterness this harvest ale has no equal

**Poor Alignment with Human-selection:** QHA = -0.3037, Full Sequence Prediction = 0.928

> like having a bowl of honey nut of malt for breakfast this beer makes me as happy as a pearly going all from tree to tree the fat is actually pretty interesting especially when compared to most brown ales its aromas and flavor are a perfectly balanced blend of honey hazelnut toast walnut and a tad smidgen of vanilla the mouthfeel is light and smooth not hoppy in the slightest quite sweet too damn drinkable for our own good

Figure S18: Beer reviews (aroma) in which human-selected sentences (underlined) are aligned well (top) and poorly (bottom) with predictive model. Fraction of SIS in the human sentences corresponds accordingly. In the bottom example (poor alignment between human-selection and predictive model), our procedure has surfaced a case where the LSTM has learned features that diverge from what a human would expect (and may suggest overfitting).
Table S3: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have positive aroma by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>smell amazing wonderful</td>
<td>2</td>
<td>nice wonderful nose</td>
<td>2</td>
<td>wonderful amazing</td>
<td>2</td>
<td>amazing amazing</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>grapefruit mango</td>
<td>2</td>
<td>pineapple grapefruit</td>
<td>1</td>
<td>hops grapefruit</td>
<td>1</td>
<td>mango pineapple</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>pineapple</td>
<td></td>
<td>grapefruit</td>
<td></td>
<td>pineapple floyds</td>
<td></td>
<td>incredible</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>nice smell citrus</td>
<td>1</td>
<td>smell great complex</td>
<td>1</td>
<td>nice smell hop</td>
<td>1</td>
<td>love</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>fresh great fantastic</td>
<td>1</td>
<td>ripe taste</td>
<td>1</td>
<td>smell pine taste</td>
<td></td>
<td>everyone</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>love</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>awesome bounds</td>
<td>1</td>
<td>rich great</td>
<td>1</td>
<td>fantastic</td>
<td>1</td>
<td>awesome nailed</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>complex</td>
<td></td>
<td>cherries</td>
<td></td>
<td>nailed</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ripe taste</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>creme brulee brulee</td>
<td>3</td>
<td>awesome grapefruit</td>
<td>1</td>
<td>incredible</td>
<td>1</td>
<td>chocolate aroma</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>creme brulee</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>chocolate cinnamon</td>
<td>1</td>
<td>dose oak</td>
<td>1</td>
<td>vanilla figs oak</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vanilla oak</td>
<td></td>
<td>acidic</td>
<td></td>
<td>thinner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>rice rice</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S4: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have negative aroma by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>awful</td>
<td>15</td>
<td>skunky</td>
<td>9</td>
<td>skunky t</td>
<td>7</td>
<td>skunky taste</td>
<td>6</td>
</tr>
<tr>
<td>C2</td>
<td>garbage</td>
<td>3</td>
<td>taste</td>
<td>1</td>
<td>garbage avoid</td>
<td>1</td>
<td>garbage rice</td>
<td>1</td>
</tr>
<tr>
<td>C3</td>
<td>vomit</td>
<td>16</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C4</td>
<td>gross rotten</td>
<td>1</td>
<td>rotten</td>
<td>1</td>
<td>awkward rotten</td>
<td>1</td>
<td>rotten offputting</td>
<td>1</td>
</tr>
<tr>
<td>C5</td>
<td>rancid horrid</td>
<td>1</td>
<td>rancid t</td>
<td>1</td>
<td>rancid</td>
<td>1</td>
<td>rancid avoid</td>
<td>1</td>
</tr>
<tr>
<td>C6</td>
<td>rice t rice</td>
<td>2</td>
<td>rice</td>
<td>1</td>
<td>rice tasteless</td>
<td>1</td>
<td>budweiser rice</td>
<td>1</td>
</tr>
</tbody>
</table>
Note that a new set of embeddings was learned with the CNN. As with the LSTM model, we use Adam (Kingma and Ba, 2015) to minimize MSE on the training set. For the aroma aspect, this CNN achieves 0.016 (0.850), 0.025 (0.748), 0.026 (0.741), 0.014 (0.662) MSE (and Pearson $\rho$) on the Train, Validation, Test, and Annotation sets, respectively. We note that this performance is very similar to that from the LSTM (see Table S1).

We apply our procedure to extract the SIS-collection from all applicable test examples using the CNN, as in Section 4.1. Figure 10a shows the predictions from one model (LSTM or CNN) when fed input examples that are SIS extracted with respect to the other model (for reviews predicted to have positive sentiment toward the aroma aspect). For example, in Figure 10a, “CNN SIS Preds by LSTM” refers to predictions made by the LSTM on the set of sufficient input subsets produced by applying our SIScollection procedure on all examples $x \in X_{test}$ for which $f_{CNN}(x) \geq \tau_+$. Since the word embeddings are model-specific, we embed each SIS using the embeddings of the model making the prediction (note that while the embeddings are different, the vocabulary is the same across the models).

In Table 2, we show five example clusters (and cluster composition) resulting from clustering the combined set of all sufficient input subsets extracted by the LSTM and CNN on reviews in the test set for which a model predicts positive sentiment toward the aroma aspect. The complete clustering on reviews receiving positive sentiment predictions is shown in Table S5 and in Table S6 for reviews receiving negative sentiment predictions.

Table S5: Joint clustering of the SIS extracted from beer reviews predicted to have positive aroma by LSTM or CNN model. Frequency indicates the number of occurrences of the SIS in the cluster. Percentages quantify SIS per cluster from the LSTM. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$ (LSTM: 20%)</td>
<td>rich chocolate</td>
<td>13</td>
<td>very rich</td>
<td>9</td>
<td>chocolate complex</td>
<td>5</td>
<td>smells rich</td>
<td>4</td>
</tr>
<tr>
<td>$C_2$ (LSTM: 21%)</td>
<td>great</td>
<td>248</td>
<td>amazing</td>
<td>119</td>
<td>wonderful</td>
<td>112</td>
<td>fantastic</td>
<td>75</td>
</tr>
<tr>
<td>$C_3$ (LSTM: 47%)</td>
<td>best smelling</td>
<td>23</td>
<td>pineapple mango</td>
<td>6</td>
<td>mango pineapple</td>
<td>6</td>
<td>pineapple grapefruit</td>
<td>5</td>
</tr>
<tr>
<td>$C_4$ (LSTM: 5%)</td>
<td>excellent</td>
<td>42</td>
<td>excellent</td>
<td>1</td>
<td>excellent</td>
<td>1</td>
<td>phenomenal</td>
<td>-</td>
</tr>
<tr>
<td>$C_5$ (LSTM: 33%)</td>
<td>oak chocolate</td>
<td>2</td>
<td>raisins oak</td>
<td>1</td>
<td>chocolate oak</td>
<td>1</td>
<td>raisins chocolate</td>
<td>1</td>
</tr>
<tr>
<td>$C_6$ (LSTM: 5%)</td>
<td>goodness</td>
<td>19</td>
<td>watering goodness</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_7$ (LSTM: 24%)</td>
<td>pumpkin pie</td>
<td>25</td>
<td>huge pumpkin</td>
<td>1</td>
<td>aroma perfect</td>
<td>1</td>
<td>smell pumpkin</td>
<td>1</td>
</tr>
<tr>
<td>$C_8$ (LSTM: 5%)</td>
<td>jd</td>
<td>13</td>
<td>tremendous</td>
<td>8</td>
<td>tremendous jd</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_9$ (LSTM: 40%)</td>
<td>brulee</td>
<td>14</td>
<td>creme brulee</td>
<td>3</td>
<td>creme creme</td>
<td>1</td>
<td>creme brulee</td>
<td>1</td>
</tr>
<tr>
<td>$C_{10}$ (LSTM: 0%)</td>
<td>s wow</td>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_{11}$ (LSTM: 0%)</td>
<td>delicious</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_{12}$ (LSTM: 0%)</td>
<td>very nice</td>
<td>23</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_{13}$ (LSTM: 70%)</td>
<td>complex aroma</td>
<td>5</td>
<td>aroma complex peaches complex</td>
<td>1</td>
<td>aroma complex interesting cherries</td>
<td>1</td>
<td>aroma complex</td>
<td>1</td>
</tr>
</tbody>
</table>

---

7For experiments involving clustering and/or comparing different models, we use examples drawn from the Test fold (instead of Annotation fold, see Table S1) to consider a larger number of examples.
Table S6: Joint clustering of the SIS extracted from beer reviews predicted to have negative aroma by LSTM or CNN model. Frequency indicates the number of occurrences of the SIS in the cluster. Percentages quantify SIS per cluster from the LSTM. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁ (LSTM: 29%)</td>
<td>not</td>
<td>247</td>
<td>no</td>
<td>105</td>
<td>bad</td>
<td>104</td>
<td>macro</td>
<td>94</td>
</tr>
<tr>
<td>C₂ (LSTM: 100%)</td>
<td>gross rotten</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₃ (LSTM: 100%)</td>
<td>rotten garbage</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₄ (LSTM: 62%)</td>
<td>vomit</td>
<td>26</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₅ (LSTM: 21%)</td>
<td>budweiser</td>
<td>22</td>
<td>sewage budweiser</td>
<td>1</td>
<td>metal budweiser</td>
<td>1</td>
<td>budweiser</td>
<td>1</td>
</tr>
<tr>
<td>C₆ (LSTM: 100%)</td>
<td>garbage rice</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>budweiser</td>
<td>1</td>
</tr>
<tr>
<td>C₇ (LSTM: 3%)</td>
<td>n’t</td>
<td>19</td>
<td>adjuncts</td>
<td>14</td>
<td>n’t adjuncts</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₈ (LSTM: 0%)</td>
<td>faint</td>
<td>82</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₉ (LSTM: 0%)</td>
<td>adjunct</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
S4.7 Results for Appearance and Palate aspects

For posterity, we include results here from repeating the analysis in our paper for the two other non-aroma aspects measured in the beer reviews data: appearance and palate.

![Graph S19](image1.png)

Figure S19: Change in appearance prediction \((f(x \setminus \{i\}) - f(x))\) after masking a randomly chosen word with mean imputation or hot-deck imputation. 10,000 replacements were sampled from the appearance beer reviews training set.

![Graph S20](image2.png)

![Graph S21](image3.png)

Figure S20: Predictive distribution on the annotation set (held-out) using the LSTM model for appearance. Vertical lines indicate decision thresholds \((\tau_+ = 0.85, \tau_- = 0.45)\) selected for \textit{SIScollection}.

Figure S21: Number of sufficient input subsets for appearance identified by \textit{SIScollection} per example.
Figure S22: Length of rationales for appearance prediction

Figure S23: Importance of individual features for appearance prediction in beer review

Table S7: Statistics for rationale length and feature importance in appearance prediction. For rationale length, median and max indicate percentage of input text in the rationale. For marginal perturbed feature importance, we indicate the median importance of features in rationales and features from the other (non-rationale) text. $p$-values are computed using a Wilcoxon rank-sum test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rationale Length (% of text)</th>
<th>Marginal Perturbed Feature Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med.</td>
<td>Max</td>
</tr>
<tr>
<td>SIS</td>
<td>2.6%</td>
<td>10.6%</td>
</tr>
<tr>
<td>Suff. IG</td>
<td>3.7%</td>
<td>89.3%</td>
</tr>
<tr>
<td>Suff. LIME</td>
<td>3.7%</td>
<td>98.2%</td>
</tr>
<tr>
<td>Suff. Perturb.</td>
<td>3.0%</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

Figure S24: QHS vs. fraction of SIS in human rationale for appearance prediction

Figure S25: Prediction on rationales only vs. rationale length for various methods in positive sentiment examples for appearance. The threshold for sufficiency was $\tau_+ = 0.85$. 
S4.7 Results for Appearance and Palate aspects

Figure S26: Prediction history on remaining (unmasked) text at each step of the BackSelect procedure, for examples where appearance sentiment is predicted.

Table S8: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have positive appearance by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>beautiful</td>
<td>376</td>
<td>nitro</td>
<td>51</td>
<td>looks great</td>
<td>38</td>
<td>great looking</td>
<td>32</td>
</tr>
<tr>
<td>C₂</td>
<td>gorgeous</td>
<td>83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₃</td>
<td>beautifully</td>
<td>7</td>
<td>absolutely</td>
<td>2</td>
<td>beautifully</td>
<td>1</td>
<td>beautifully bands</td>
<td>1</td>
</tr>
<tr>
<td>C₄</td>
<td>brilliant</td>
<td>5</td>
<td>brilliant slowly</td>
<td>1</td>
<td>wonderfully</td>
<td>1</td>
<td>appearance brilliant</td>
<td>1</td>
</tr>
<tr>
<td>C₅</td>
<td>lovely looking</td>
<td>3</td>
<td>black lovely</td>
<td>3</td>
<td>impressively</td>
<td>1</td>
<td>lovely crystal</td>
<td>1</td>
</tr>
</tbody>
</table>

Table S9: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have negative appearance by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>piss</td>
<td>46</td>
<td>zero</td>
<td>38</td>
<td>water</td>
<td>37</td>
<td>water</td>
<td>27</td>
</tr>
<tr>
<td>C₂</td>
<td>unappealing</td>
<td>12</td>
<td>floaties</td>
<td>12</td>
<td>floaties unappealing</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C₃</td>
<td>ugly</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table S10: Statistics for rationale length and feature importance in palate prediction. For rationale length, median and max indicate percentage of input text in the rationale. For marginal perturbed feature importance, we indicate the median importance of features in rationales and features from the other (non-rationale) text. *p*-values are computed using a Wilcoxon rank-sum test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rationale Length (% of text)</th>
<th>Marginal Perturbed Feature Importance</th>
<th>p (vs. SIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Med.</td>
<td>Max</td>
<td>p (vs. SIS)</td>
</tr>
<tr>
<td>SIS</td>
<td>2.4%</td>
<td>13.7%</td>
<td>–</td>
</tr>
<tr>
<td>Suff. IG</td>
<td>3.2%</td>
<td>56.1%</td>
<td>2e-06</td>
</tr>
<tr>
<td>Suff. LIME</td>
<td>3.0%</td>
<td>57.0%</td>
<td>7e-06</td>
</tr>
<tr>
<td>Suff. Perturb.</td>
<td>2.8%</td>
<td><strong>11.8%</strong></td>
<td>3e-03</td>
</tr>
</tbody>
</table>
Figure S27: Change in palate prediction \( f(x \setminus \{i\}) - f(x) \) after masking a randomly chosen word with mean imputation or hot-deck imputation. 10,000 replacements were sampled from the palate beer reviews training set.

Figure S28: Predictive distribution on the annotation set (held-out) using the LSTM model for palate. Vertical lines indicate decision thresholds (\( \tau_+ = 0.85 \), \( \tau_- = 0.45 \)) selected for SIScollection.

Figure S29: Number of sufficient input subsets for palate identified by SIScollection per example.

Figure S30: Length of rationales for palate prediction

Figure S31: Importance of individual features in beer review palate rationales
S4.7 Results for Appearance and Palate aspects

Figure S32: QHS vs. fraction of SIS in human rationale for palate prediction

Figure S33: Prediction on rationales only vs. rationale length for various methods in positive sentiment examples for palate. The threshold for sufficiency was $\tau_+ = 0.85$.

Figure S34: Prediction history on remaining (unmasked) text at each step of the BackSelect procedure, for examples where palate sentiment is predicted.

Table S11: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have positive palate by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster. Dashes are used in clusters with under 4 unique SIS.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>smooth creamy</td>
<td>27</td>
<td>silky smooth</td>
<td>20</td>
<td>mouthfeel perfect</td>
<td>16</td>
<td>creamy perfect</td>
<td>12</td>
</tr>
<tr>
<td>$C_2$</td>
<td>mouthfeel</td>
<td>6</td>
<td>exceptional</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_3$</td>
<td>perfect</td>
<td>50</td>
<td>perfect</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_4$</td>
<td>smooth velvety</td>
<td>6</td>
<td>velvety smooth</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_5$</td>
<td>silk</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_6$</td>
<td>smooth perfect</td>
<td>8</td>
<td>mouth smooth</td>
<td>1</td>
<td>perfect smooth</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_7$</td>
<td>perfect great</td>
<td>5</td>
<td>great perfect</td>
<td>2</td>
<td>feels perfect</td>
<td>2</td>
<td>perfect feels great</td>
<td>1</td>
</tr>
</tbody>
</table>
Table S12: All clusters of sufficient input subsets extracted from reviews from the test set predicted to have negative palate by the LSTM. Frequency indicates the number of occurrences of the SIS in the cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SIS #1</th>
<th>Freq.</th>
<th>SIS #2</th>
<th>Freq.</th>
<th>SIS #3</th>
<th>Freq.</th>
<th>SIS #4</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>overcarbonated</td>
<td>12</td>
<td>mouthfeel</td>
<td>3</td>
<td>way overcarbonated</td>
<td>1</td>
<td>overcarbonated</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>overcarbonated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₂</td>
<td>watery</td>
<td>302</td>
<td>thin</td>
<td>238</td>
<td>flat</td>
<td>118</td>
<td>mouthfeel thin</td>
<td>33</td>
</tr>
<tr>
<td>C₃</td>
<td>too carbonation</td>
<td>1</td>
<td>too carbonation</td>
<td>1</td>
<td>mouthfeel odd too</td>
<td>1</td>
<td>too carbonated</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>masks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₄</td>
<td>lack carbonation</td>
<td>4</td>
<td>carbonation lack</td>
<td>4</td>
<td>carbonation hurts</td>
<td>2</td>
<td>issue lack hurts</td>
<td>1</td>
</tr>
</tbody>
</table>
Additional References for the Supplementary Information


