Rationalization through Concepts

Diego Antognini and **Boi Faltings**
École Polytechnique Fédérale de Lausanne, Switzerland
firstname.lastname@epfl.ch

Abstract

Automated predictions require explanations to be interpretable by humans. One type of explanation is a rationale, i.e., a selection of input features such as relevant text snippets from which the model computes the outcome. However, a single overall selection does not provide a complete explanation, e.g., weighing several aspects for decisions. To this end, we present a novel self-interpretable model called ConRAT. Inspired by how human explanations for high-level decisions are often based on key concepts, ConRAT extracts a set of text snippets as concepts and infers which ones are described in the document. Then, it explains the outcome with a linear aggregation of concepts. Two regularizers drive ConRAT to build interpretable concepts. In addition, we propose two techniques to boost the rationale and predictive performance further. Experiments on both single- and multi-aspect sentiment classification tasks show that ConRAT is the first to generate concepts that align with human rationalization while using only the overall label. Further, it outperforms state-of-the-art methods trained on each aspect label independently.

1 Introduction

Neural models have become the standard for many tasks, owing to their large performance gains. However, their adoption in decision-critical fields is more limited because of their lack of interpretability, particularly with textual data.

One of the simplest means of explaining predictions of complex models is by selecting relevant input features. Attention mechanisms (Bahdanau et al., 2015) model the selection using a conditional importance distribution over the inputs, but the resulting explanations are noisy (Jain and Wallace, 2019; Pruthi et al., 2020). Multi-head attention (Vaswani et al., 2017) extends attention mechanisms to attend information from different perspectives jointly. However, no explicit mechanisms guarantee a logical connection between different views (Voita et al., 2019; Kovaleva et al., 2019). Another line of research includes rationale generation methods (Lei et al., 2016; Chang et al., 2020; Antognini et al., 2021b). If the selected text input features are short and concise – called a rationale – and suffice on their own to yield the prediction, it can potentially be understood and verified against domain knowledge (Chang et al., 2019).

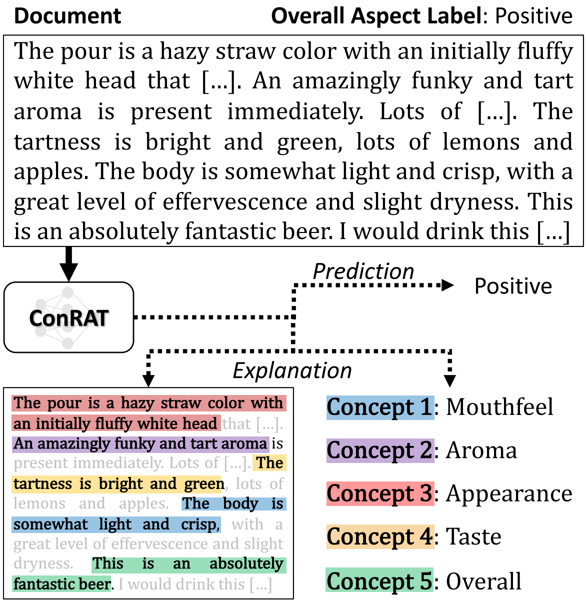

Document
Overall Aspect Label: Positive
The pour is a hazy straw color with an initially fluffy white head that [...]. An amazingly funky and tart aroma is present immediately. Lots of [...]. The tartness is bright and green, lots of lemons and apples. The body is somewhat light and crisp, with a great level of effervescence and slight dryness. This is an absolutely fantastic beer. I would drink this [...]
ConRAT
Prediction
Positive
Explanation
The pour is a hazy straw color with an initially fluffy white head that [...]. An amazingly funky and tart aroma is present immediately. Lots of [...]. The tartness is bright and green, lots of lemons and apples. The body is somewhat light and crisp, with a great level of effervescence and slight dryness. This is an absolutely fantastic beer. I would drink this [...]
Concept 1: Mouthfeel
Concept 2: Aroma
Concept 3: Appearance
Concept 4: Taste
Concept 5: Overall

Figure 1: An illustration of ConRAT. Given a beer review, ConRAT identifies five excerpts that relate to particular concepts of beers (i.e., the explanation), depicted in color, from which it computes the outcome.

The key motivation for this work arises from the limitations of rationales. Rationalization models strive for one overall selection to explain the outcome by maximizing the mutual information between the rationale and the label. However, useful rationales can be multi-faceted, where each facet relates to a particular "concept" (see Figure 1). For example, users typically justify their opinions of a product by weighing explanations: one for each aspect they care about (Musat and Faltings, 2015).

Inspired by how human reasoning comprises concept-based thinking (Armstrong et al., 1983; Tenenbaum, 1999), we aim to discover, in an unsupervised manner, a set of concepts to explain the outcome with a weighted average, similar to multi-head attention. In this work, we relate concepts to semantically meaningful and consistent excerpts across multiple texts. Unlike topic modeling, where documents are described by a set of latent topics comprising word distributions, our latent concepts relate to text snippets that are relevant for the prediction.

Another motivation for this study is to generate interpretable concepts. The explanation of an outcome should rely on concepts that satisfy the desiderata introduced in Alvarez-Melis and Jaakkola (2018). They should 1. preserve relevant information, 2. not overlap with each other and be diverse, and 3. be human-understandable. Figure 1 shows an example of concepts in the beer domain.

In this work, we present a novel self-explaining neural model: the concept-based rationalizer (ConRAT) (see Figure 1 and 2). Our new rationalization scheme first identifies a set of concepts in a document and then decides which ones are currently described (binary selection). ConRAT explains the prediction with a linear aggregation of concepts. The model is trained end-to-end, and the concepts are learned in an unsupervised manner. In addition, we design two regularizers that guide ConRAT to induce interpretable concepts and propose two optional techniques, knowledge distillation and concept pruning, in order to boost the performance further.

We evaluate ConRAT on both single- and multi-aspect sentiment classification with up to five target labels. Upon training ConRAT only on the overall aspect, the results show that ConRAT generates concepts that are relevant, diverse, and non-overlapping, and they also recover human-defined concepts. Furthermore, our model significantly outperforms strong supervised baseline models in terms of predictive and explanation performance.

2 Related Work

Developing interpretable models is of considerable interest to the broader research community. Researchers have investigated many approaches to improve the interpretability of neural networks.

2.1 Interpretability.

The first line of research aims at providing post-hoc explanations of an already trained model. For example, gradient and perturbation-based methods attribute the decision to important input features (Ribeiro et al., 2016; Sundararajan et al., 2017; Lundberg and Lee, 2017; Shrikumar et al., 2017). Other studies identified the causal relationships between input-output pairs (Alvarez-Melis and Jaakkola, 2017; Goyal et al., 2019). In contrast, our model is inherently interpretable as it directly produces the prediction with an explanation.

Another line of research has developed interpretable models. Quint et al. (2018) extended a variational auto-encoder with a differentiable decision tree. Alaniz and Akata (2019) proposed an explainable observer-classifier framework whose predictions can be exposed as a binary tree. However, these methods have been designed for images only, while our work focuses on text input.

The works most relevant to ours relate to interpretable models from the rationalization field (Lei et al., 2016; Bastings et al., 2019; Yu et al., 2019; Chang et al., 2020; Jain et al., 2020; Paranjape et al., 2020). These methods justify their predictions by selecting rationales (i.e., relevant tokens in the input text). However, they are limited to explain only the prediction with mostly one text span and rely on the assumption that the data have low internal correlations (Antognini et al., 2021b). Chang et al. (2019) extended previous methods to extract an additional rationale in order to counter the prediction. In our work, ConRAT produces multi-faceted rationales and explains the prediction through a linear aggregation of the extracted concepts. However, if we set the number of concepts to one, ConRAT reduces to a special case of a rationale model.

2.2 Explanations through Concepts.

Researchers have proposed multiple approaches for concept-based explanations. Kim et al. (2018) designed a post-hoc technique to learn concept activation vectors by relying on human annotations that characterize concepts of interest. Similarly, Bau et al. (2017); Zhou et al. (2018) generated visual explanations for a classifier. Our concepts are learned in an unsupervised manner and not defined a priori.

Few studies have learned concepts on images in an unsupervised fashion. Li et al. (2018) explained predictions based on the similarity of the input to "prototypes" learned during training. Alvarez-Melis and Jaakkola (2018) used an auto-encoder to extract relevant concepts and explain the prediction. Ghorbani et al. (2019) designed an unsupervised concept discovery method to explain trained models. Koh et al. (2020) employed the discovered concepts to predict the target label. Our work's key difference is that we focus on text data, while all these methods treat only image inputs.

To the best of our knowledge, Bouchacourt and Denoyer (2019) is the only study that has proposed a self-interpretable concept-based model for text data using reinforcement learning. It computes the predictions and provides an explanation in terms of the presence or absence of concepts in the input (i.e., text excerpts of variable lengths). However, their method achieves poor overall performance. In addition, it is unclear whether the discovered concepts are interpretable. Conversely, ConRAT is differentiable, clearly outperforms strong models in terms of predictive and explanation performance, and it infers relevant, diverse, non-overlapping, and human-understandable concepts.

2.3 Topic Modeling.

Topic models, such as latent Dirichlet allocation (Blei et al., 2003), describe documents with a mixture of latent topics. Each topic represents a word distribution. Some studies combined topic models with recurrent neural models (Dieng et al., 2017; Zaheer et al., 2017). However, the goal of these generative models and the topics remains different than this work's. We aim to build a self-interpretable model that predicts and explains the outcome with latent concepts.

3 Concept-based Rationalizer (ConRAT)

Figure 2 depicts the architecture of our proposed self-explaining model: the Concept-based Rationalizer (ConRAT). Let X be a random variable representing a document composed of T words $(x_1, x_2, \dots, x_T)$, y the ground-truth label, and K the desired numbers of concepts.[1] Given a document X and a label y, our goal is to explain the prediction $\hat{y}$ by finding a set of K concepts C_1, $\dots, C_K$ that are masked versions of X. ConRAT learns concepts by maximizing the mutual information between $\boldsymbol{C}$ and y. We guide ConRAT to create separable and consistent concepts via two regularizers to make them human-understandable.

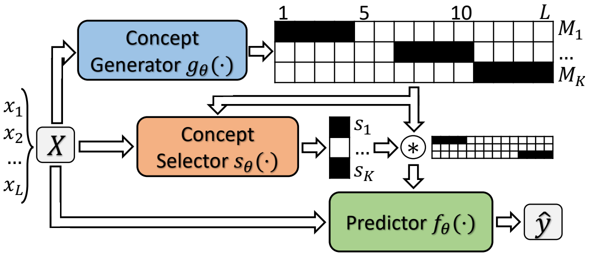

1
5
10
L
Concept Generator $g_\theta(\cdot)$
M_1
...
M_K
x_1
x_2
...
x_L
X
Concept Selector $s_\theta(\cdot)$
s_1
...
s_K
$*$
Predictor $f_\theta(\cdot)$
$\hat{y}$

Figure 2: The proposed self-explaining model ConRAT. The model predicts and explains $\hat{y}$. Given a document X, the concept generator produces one binary mask per concept. The concept selector decides which concepts are present in the input. The predictor aggregates each selected concept's prediction to compute $\hat{y}$.

3.1 Model Overview

ConRAT is divided into three submodels: a **Concept Generator** $g_\theta(\cdot)$, which finds the concepts $C_1, \dots, C_K$; a **Concept Selector** $s_\theta(\cdot)$, which detects whether a concept C_k is present or absent (i.e., $s_k \in \{1, 0\}$) in the input X; and a **Predictor** $f_\theta(\cdot)$, which predicts the outcome $\hat{y}$ based on the concepts $\boldsymbol{C}$ and their presence scores S.

3.1.1 Concept Generation

Inspired by the selective rationalization field (Lei et al., 2016), we define "concept" as a sequence of consecutive words in the input text. Previous studies extracted only one concept C_1 that is sufficient to explain the target variable y. In our work, a major difference is that we aim to find K concepts $C_1, \cdots, C_K$ that represent different topics or aspects and altogether explain the target variable y. We interpret the model as being linear in the concepts rather than depending on one overall selection of word. More formally, we define a concept as follows:

$$C_k = M_k \odot X, \tag{1}$$

where $M_k \in \mathbb{S}$ denotes a binary mask, $\mathbb{S}$ is a subset of $\mathbb{Z}_2^T$ with some constraints (introduced in Section 3.2), and $\odot$ is the element-wise multiplication of two vectors.

We parametrize the binary masks $\boldsymbol{M} \in \mathbb{Z}_2^{K \times T}$ with the concept generator model $g_\theta(\cdot)$, based on a bi-directional recurrent neural network. Following previous rationalization research (Yu et al.,

[1] Our method is easily adapted for regression problems.

2019; Chang et al., 2020), we force $g_\theta(\cdot)$ to select one chunk of text per concept with a prespecified length $\ell \in [1, T]$.[2] Instead of predicting the mask M_k directly, $g_\theta(\cdot)$ produces a score for each position t. Then, it samples the start position t^*_k of the chunk for each C_k using the straight-through Gumbel-Softmax (Maddison et al., 2017; Jang et al., 2017). Finally, we compute M_k as follows:

$$\begin{aligned} &T^* \sim Gumbel(g_\theta(X)), \\ &M_{k,t} = \mathbb{1}\left[t \in [t^*_k, min(t^*_k + \ell - 1, T)]\right], \end{aligned} \tag{2}$$

where $\mathbb{1}$ denotes the indicator function. Although the equation is not differentiable, we can employ the straight-through technique (Bengio et al., 2013) and approximate it with the gradient of a causal convolution and a convolution kernel of an all-one vector of length ℓ.

3.1.2 Concept Selection

A key objective of ConRAT is to produce semantically consistent and separable concepts. So far, the generator $g_\theta(\cdot)$ generates K concepts for any input document. However, some documents might mention only a subset of those. Thus, the goal of the concept selector model $s_\theta(\cdot)$ is to enable ConRAT to ignore absent concepts.

Specifically, for each concept C_k, the model first computes a concept representation H_{C_k} using a standard attention mechanism (Bahdanau et al., 2015) (the tokens whose $M_{k,t} = 0$ are masked out). Then, we take the dot product of H_{C_k} with a weight vector, followed by a sigmoid activation function to induce the log-probabilities of a relaxed Bernoulli distribution (Jang et al., 2017). Finally, we sample the presence score $s_k \in \{0, 1\}$ of each concept independently:

$$S \sim RelaxedBernoulli(s_\theta(X, \boldsymbol{M})). \tag{3}$$

3.1.3 Prediction

As inputs, the predictor $f_\theta(\cdot)$ takes the document X, the masks $\boldsymbol{M}$, and the presence scores S for all concepts. First, we extract the concepts, which are masked versions of X. Differently than in Equation 1, the concepts are ignored if $s_k = 0$:

$$C_k = (M_k * s_k) \odot X. \tag{4}$$

Second, the model produces the hidden representation h'_{C_k} with another recurrent neural network, followed by a LeakyReLU activation function (Xu et al., 2015). Then, it computes the logits of y by applying a linear projection for each concept:

$$P_k = W h'_{C_k} + b, \tag{5}$$

where W and b are the projection parameters. Finally, f_θ computes the final outcome as follows:

$$p(y|C, \boldsymbol{M}, X) = softmax(\sum_{k=1}^{K} \alpha_k P_k s_k), \tag{6}$$

where α_k are model parameters that can be interpreted as the degree to which a particular concept contributes to the final prediction.

3.2 Unsupervised Discovery of Concepts

The above formulations integrate the explanation into the outcome computation. However, M_k is by definition faithful to the model's inner workings but not comprehensible for the end-user. Following Alvarez-Melis and Jaakkola (2018), we aim the concepts to follow three desiderata:1. **Fidelity**: they should preserve relevant information, 2. **Diversity**: they should be non-overlapping and diverse, and 3. **Grounding**: they should have an immediate human-understandable interpretations.

The hard constraint in Equation 2 naturally enforces the grounding by forcing the concept to be a sequence of ℓ words. For the fidelity, it is partly integrated in ConRAT by the prediction loss, which is the cross-entropy between the ground-truth label y and the prediction $\hat{y}$: $\mathcal{L}_{pred} = CE(\hat{y}, y)$. Recall that the concepts are substitutes of the input that are sufficient for the prediction. We emphasize the word "partly" because nothing prevents ConRAT from picking up spurious correlations.

We propose two regularizers to encourage ConRAT in finding non-overlapping, relevant, and dissimilar concepts. The first favors the orthogonality of concepts by penalizing redundant rows in $\boldsymbol{M}$:

$$\mathcal{L}_{overlap} = ||\boldsymbol{M}\boldsymbol{M}^T - \ell \cdot \mathbb{1}\,||^2_F, \tag{7}$$

where $||\cdot||_F$ stands for the Frobenius norm of a matrix, $\mathbb{1}$ denotes the identity matrix, and ℓ the pre-specified concept length. However, $\mathcal{L}_{overlap}$ alone does not prevent ConRAT from learning little relevant concepts. Therefore, we propose a second regularizer to encourage fidelity and diversity

[2] In early experiments, we relaxed the length constraint and generated instead K differentiable masks with continuity regularizers. However, this variant produced majorly inferior results. We hypothesize that there are too many constraints to optimize with only the target label as a strong signal.

by minimizing the cosine similarity between the concept representations H_{C_k} (see Section 3.1.2):

$$\mathcal{L}_{div} = \frac{1}{K}\frac{1}{K-1}\sum_{\substack{k_1,k_2=1\\k_1\neq k_2}}^{K} cos(H_{C_{k_1}}, H_{C_{k_2}}). \quad (8)$$

In both regularizers, we do not consider the presence scores S because a model could always select only one concept; this strategy is not optimal and reduces to a special case of rationale models (i.e., S would become a one-hot vector).

To summarize, the concepts are learned in an unsupervised manner and align with the three desiderata mentioned above: diversity is achieved with $\mathcal{L}_{overlap}$ and $\mathcal{L}_{div}$; fidelity is enforced by $\mathcal{L}_{pred}$ and $\mathcal{L}_{div}$, and the hard constraint in Equation 2 ensures the grounding. Finally, we train ConRAT end-to-end and minimize the loss jointly $\mathcal{L} = \mathcal{L}_{pred} + \lambda_O\mathcal{L}_{overlap} + \lambda_D\mathcal{L}_{div}$, where λ_O and λ_D control the impact of each regularizer.

3.3 Improving Overall Performance Further

The purpose of self-explaining models is to compute outcomes while being more interpretable. However, one key point is to achieve predictive performance comparable to that of black-box models. We propose two techniques to further improve both interpretability and performance; however, ConRAT does not require these techniques to outperform other methods, as we will see later.

Knowledge Distillation. We can train ConRAT not only via the information provided by the true labels but also by observing how a teacher model behaves (Hinton et al., 2015). In that case, we introduce the teacher model $T_\theta(\cdot)$, which is a simple recurrent neural network similar to the predictor f_θ. It is trained one the same data, but it uses the whole input X instead of subsets selected by each C_k. The overall training loss becomes $\mathcal{L} = \mathcal{L}_{pred} + \lambda_O\mathcal{L}_{overlap} + \lambda_D\mathcal{L}_{div} + \lambda_T(\hat{y}_{T_\theta} - \hat{y}_{f_\theta})^2$.

Pruning Concepts. Depending on the number of concepts and the pre-specified length, the total number of selected words can be close to or higher than the document length.[3] In practice, it is hard to extract meaningful concepts in such settings. To alleviate this problem, we propose to prune concepts at inference and select the top-k concepts that overlap the least with the others. More specifically, we compute the overlap as follows: for each sample in the validation set, we measure the average overlap ratio between M_{k_1} and M_{k_2} for each concept-pair $(C_{k_1}, C_{k_2}), k_1 \neq k_2$. Then, we select the top-k concepts whose scores are the lowest. Finally, to compute the new prediction $\hat{y}$, we update $s_k = 1$ if C_k is in the top-k or $s_k = 0$ otherwise.

Dataset	Amazon	Beer
# Reviews	$24,000$	$60,000$
Split Train/Val/Test	20k/2k/2k	50k/5k/5k
# Annotations	471	994
# Human Aspects	1	5
# Words per review	224 ± 125	184 ± 58

Table 1: Statistics of the review datasets.

4 Experiments

4.1 Datasets

We evaluate the quantitative performance of ConRAT using two binary classification datasets. The first one is the single-aspect Amazon Electronics dataset (Ni et al., 2019). We followed the filtering process in Chang et al. (2019) to keep only the reviews that contain evidence for both positive and negative sentiments. Specifically, we considered the first 50 tokens after the words "pros:" and "cons:" as the rationale annotations for the positive and negative labels, respectively. We randomly picked 24,000 balanced samples with ratings of four and above or two and below.

The second dataset comprises the multi-aspect beer reviews (McAuley et al., 2012) used in the field of rationalization (Lei et al., 2016; Yu et al., 2019). Each review describes various beer aspects: Appearance, Aroma, Palate, Taste, and Overall; users also provided a five-star rating for each aspect. However, we only use the overall rating for ConRAT. The dataset includes 994 beer reviews with sentence-level aspect annotations. Following the evaluation protocol in Bao et al. (2018); Chang et al. (2020), we binarized the ratings ≤ 2 as negative and ≥ 3 as positive. We sampled 60,000 balanced examples. Our setting is more challenging than those in previous studies because we assess the performance on all aspects (instead of three) and consider all examples for the sampling (instead of de-correlated subsets), reflecting the real data distribution. Table 1 shows the data statistics.

[3] e.g., if a document contains 200 tokens and we aim to extract 10 concepts of 20 tokens, all words should be selected.

4.2 Baselines

We consider the following baselines. **RNP** is a generator-predictor framework proposed by Lei et al. (2016) for rationalizing neural prediction. The generator selects text spans as rationales, which are then fed to the classifier for the final prediction. Yu et al. (2019) introduced **RNP-3P**, which extends RNP to include the complement predictor as the third player. It maximizes the predictive accuracy from unselected words. The training consists of an adversarial game with the three players. **Intro-3P** (Yu et al., 2019) improves RNP-3P by conditioning the generator on the predicted outcome of a teacher model. **InvRAT** is a game-theoretic method that competitively rules out spurious words with strong correlations to the output. The game-theoretic approach **CAR** aims to infer a rationale and a counterfactual rationale that counters the true label. We follow Chang et al. (2020) and consider for all methods their hard constraint variant (i.e., selecting one chunk of text) with different lengths for generating rationales.

RNP-3P and Intro-3P are trained with the policy gradient (Williams, 1992). The others estimate the gradients of the rationale selections using the straight-through technique (Bengio et al., 2013).

All rationalization methods, except CAR, strive for a single overall selection ($K = 1$) to explain the outcome. For the multi-aspect dataset, we train and tune each baseline independently for each aspect. The key difference with ConRAT is that the model is only trained on the overall aspect label and infers one rationale of K concepts; the baselines are trained K times to infer one rationale of one concept.

4.3 Experimental Details

To seek fair comparisons, we try to keep a similar number of parameters across all models, and we employ the same architecture for each player (generators, predictors, and discriminators/teachers) in all models: bi-directional gated recurrent units (Chung et al., 2014) with a hidden dimension 256. We use the 100-dimensional GloVe word embeddings (Pennington et al., 2014), Adam (Kingma and Ba, 2015) as optimization method with a learning rate of 0.001. We set the convolutional neural network in the concept selector similarly to (Kim et al., 2015) with 3-, 5-, and 7-width filters and 50 feature maps per filter. For ConRAT, we set the regularizer factors as follow: $\lambda_O = 0.05$, $\lambda_D = 0.05$, and $\lambda_T = 0.5$. We use the open-source implementation for all models, and we tune them by maximizing the prediction accuracy on the dev set with 16 random searches. For reproducibility purposes, we include additional details in Appendix A.

Table 2: Accuracy and objective performance of rationales in automatic evaluation for the Amazon dataset.

		Factual			*Counter Fact.*		
Model	**Acc.**	**P**	**R**	**F**	**P**	**R**	**F**
RNP	**75.5**	32.6	18.8	23.8		–	
RNP-3P	70.0	49.4	28.4	36.0		–	
Intro-3P	75.2	22.1	12.8	16.2		–	
InvRAT	71.5	44.3	25.5	32.4		–	
ConRAT-1	**75.5**	**56.4**	**32.5**	**41.3**		–	
CAR	73.6	33.0	19.1	24.2	**44.1**	**25.4**	**32.2**
ConRAT-6	**75.4**	**50.0**	**28.8**	**36.6**	32.3	18.6	23.6
ConRAT-4	75.3	46.4	26.7	33.9	29.6	17.1	21.6
ConRAT-2	75.3	33.7	19.4	24.6	8.9	5.1	6.5

4.4 RQ 1: Can ConRAT find evidence for factual and counterfactual rationales?

We aim to validate whether ConRAT can identify the two evidences for positive and negative sentiments. We set the concept length $\ell = 30$, we compare the generated rationales with the annotations, and we report the precision, recall, and F1 score. In this experiment, no teacher is used in ConRAT.

Table 2 contains the results. The top rows contain the results when only the factual rationales are considered for the evaluation, and ConRAT-1 uses only one concept. We see that ConRAT surpasses the baselines in finding rationales that align with human annotations, and it also matches the test accuracy with the baselines. Interestingly, we note that the baselines achieving the highest accuracy underperform in finding the correct rationales.

For the factual and counterfactual rationales, CAR finds one rationale to support the outcome and another one to counter it, in an adversarial game. However, the concepts inferred by ConRAT are not guaranteed to align with the rationales as there is no explicit signal to infer counterfactual concepts. Thus, we increase the number of concepts up to six and prune ConRAT to consider only the two most dissimilar concepts (see Section 3.3).

The bottom of Table 2 show the results. With only two concepts, ConRAT-2 outperforms CAR

Table 3: Objective performance of rationales for the multi-aspect beer reviews. ConRAT only uses the overall label and ignores the other aspect labels. All baselines are trained separately on each aspect rating. **Bold** and underline denote the best and second-best results, respectively.

			Average			*Appearance*			*Aroma*			*Palate*			*Taste*			*Overall*		
	Model	**Acc.**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
$\ell = 20$	RNP	81.1	30.7	22.1	24.9	30.8	23.2	26.5	22.1	21.0	21.5	17.7	24.1	20.4	28.1	16.7	20.9	54.9	25.8	35.1
	RNP-3P	80.5	29.1	22.5	25.0	30.4	25.6	27.8	19.3	20.4	19.8	10.3	12.0	11.1	43.9	28.4	34.5	41.6	26.0	32.0
	Intro-3P	85.6	24.2	19.6	21.3	28.7	24.8	26.6	14.3	14.4	14.3	16.6	19.3	17.9	24.2	13.6	17.4	37.0	25.9	30.5
	InvRAT	82.9	41.8	31.1	34.8	54.5	45.5	49.6	26.1	27.6	26.9	22.6	25.9	24.1	46.6	27.4	34.5	59.0	29.3	39.2
	ConRAT*	**91.4**	**50.0**	**42.0**	**44.9**	**57.8**	**53.0**	**55.3**	**31.9**	**35.5**	**33.6**	**29.0**	**36.3**	**32.3**	**56.5**	**33.9**	**42.4**	**74.9**	**51.0**	**60.7**
$\ell = 10$	RNP	84.4	41.3	16.6	23.2	40.1	12.0	18.5	**33.3**	**18.7**	**24.0**	**25.1**	**17.4**	**20.6**	32.3	9.8	15.07	76.0	25.1	37.8
	RNP-3P	83.1	31.1	13.5	18.6	41.8	19.2	26.3	22.2	12.4	15.9	16.5	10.4	12.7	33.2	10.6	16.1	41.9	14.7	21.8
	Intro-3P	80.9	21.8	10.8	14.3	51.0	26.0	34.4	18.8	9.7	12.8	16.5	10.6	12.9	9.7	2.6	4.1	13.1	5.2	7.4
	InvRAT	81.9	47.1	17.8	25.5	**59.4**	26.1	**36.3**	31.3	15.5	20.8	16.4	9.6	12.1	39.1	11.6	17.9	**89.1**	26.4	40.7
	ConRAT*	**91.3**	**48.1**	**20.1**	**28.0**	51.7	**26.2**	34.8	32.6	17.4	22.7	23.0	13.8	17.3	**45.3**	**13.1**	**20.3**	88.0	**30.1**	**44.9**

* The model is only trained on the overall label and does not have access to the other ground-truth labels.

in terms of test accuracy and matches the performance for the factual rationales, but it poorly identifies counterfactual rationales. However, there is a major improvement when we increase the number of concepts and use pruning. Indeed, the word distribution of the factual and counterfactual rationales are different, hence captured with pruning. ConRAT's factual rationales are better than those of all models. The counterfactual ones get closer to those produced by CAR. We show later in Section 4.6 that pruning helps in achieving better correlation with human judgments but is not required.

4.5 RQ 2: Are concepts inferred by ConRAT consistent with human rationalization?

We investigate whether ConRAT can recover all beer aspects by using only the overall ratings. Because beer reviews are smaller in length than Amazon ones, we set the concept length ℓ to 10 and 20. We fix the number of concepts to ten and prune ConRAT to keep five. We manually map them to the closest aspect for comparison. We trained the teacher model, used in Intro-3P and ConRAT, and obtained 91.4% accuracy. More results and illustrations are available in Appendix B and C.

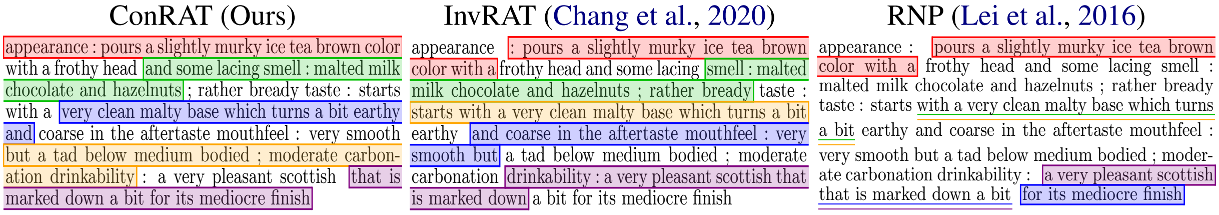
Objective Evaluation. Similar to Section 4.4, we compare the generated rationales with the human annotations on the five aspects and the average performance. The main results are shown in Table 3. On average, ConRAT achieves the best performance while trained only on the overall ratings. This shows that the generated concepts, learned in an unsupervised manner, are separable, consistent, and correlated with human judgments to a certain extent. For the concept length $\ell = 20$, ConRAT produces significant superior results for all aspects, whereas the difference with InvRAT is less pronounced for $\ell = 10$. Finally, ConRAT's concepts lead to the highest accuracy and respect the grounding desideratum, thanks to the teacher.

We hypothesize that the baselines underperform due to the high correlations among the aspect ratings. Thus, they are more prone to pick up spurious correlations between the input features and the output. By considering multiple concepts simultaneously, ConRAT reduces the impact of spurious correlations. Regarding Intro-3P and RNP-3P, both suffer from instability issues due to the policy gradient (Chang et al., 2020; Yu et al., 2019).

We visualize an example in Figure 3. We observe that ConRAT induces interpretable concepts, while the best baselines suffer from spurious correlations. By reading our concepts alone, humans will easily predict the aspect label and its polarity.

Subjective Evaluation. We conduct a human evaluation using Amazon's Mechanical Turk (details in Appendix B.2) to judge the understandability of the concepts. Following Chang et al. (2019), we sampled 100 balanced reviews from the hold-out set for each aspect, model, and concept length, resulting in 5,000 samples. We showed the examples in random order. An evaluator is presented with the concept generated by one of the five methods (unselected words are not visible). We credit a success when the evaluator guesses the true aspect label and its sentiment. We report the success rate as the performance metric. A random guess has a 10% success rate.

Figure 4 shows the main results. Similar to the objective evaluation, ConRAT reaches the



ConRAT (Ours)

appearance : pours a slightly murky ice tea brown color with a frothy head and some lacing smell : malted milk chocolate and hazelnuts ; rather bready taste : starts with a very clean malty base which turns a bit earthy and coarse in the aftertaste mouthfeel : very smooth but a tad below medium bodied ; moderate carbonation drinkability : a very pleasant scottish that is marked down a bit for its mediocre finish

InvRAT (Chang et al., 2020)

appearance : pours a slightly murky ice tea brown color with a frothy head and some lacing smell : malted milk chocolate and hazelnuts ; rather bready taste : starts with a very clean malty base which turns a bit earthy and coarse in the aftertaste mouthfeel : very smooth but a tad below medium bodied ; moderate carbonation drinkability : a very pleasant scottish that is marked down a bit for its mediocre finish

RNP (Lei et al., 2016)

appearance : pours a slightly murky ice tea brown color with a frothy head and some lacing smell : malted milk chocolate and hazelnuts ; rather bready taste : starts with a very clean malty base which turns a bit earthy and coarse in the aftertaste mouthfeel : very smooth but a tad below medium bodied ; moderate carbonation drinkability : a very pleasant scottish that is marked down a bit for its mediocre finish

Figure 3: Concepts generated (with ℓ=10) for a beer review. Underline highlights ambiguities. The color depicts the aspects: Appearance, Aroma, Palate, Taste, and Overall. **ConRAT is trained only on the overall label**.

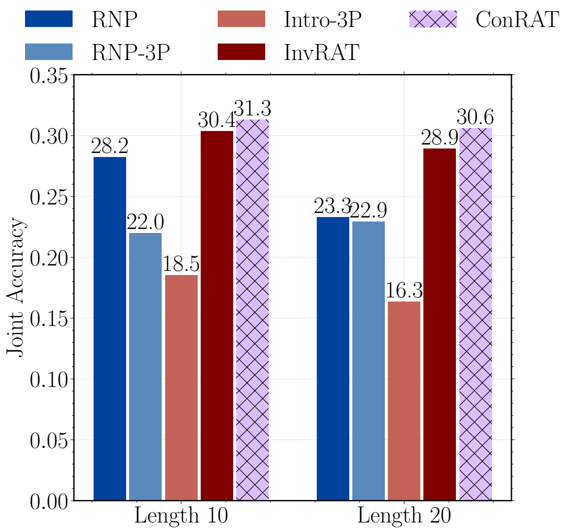


Figure 4: Subjective performance of rationales for the multi-aspect beer reviews. Evaluators need to guess both the sentiment and what aspect the concept is about, which makes random guess only 10%.

best performance, followed by InvRAT. Moreover, ConRAT only requires a single training on the overall aspect. It emphasizes that the discovered concepts satisfy the fidelity and diversity desiderata and better correlate with human judgments compared with supervised baselines.

4.6 RQ 3: How does the number of concepts K in ConRAT affect the performance?

We study the impact of the number of concepts K in ConRAT on the performance, as discussed in Section 4.5. We set the number of concepts to the number of aspects (K=5) and then increase it to K=10 and K=20. We prune ConRAT to keep only the five most dissimilar concepts (see Section 3.3).

Results are shown in Table 4. First, we observe that the performance is already better than the baselines in Table 3 with K=5. Second, when increasing K and pruning ConRAT, the performance is boosted further. However, we remark that the interpretability of the concepts follows a bell curve and significantly decreases when K=20. One potential reason is that we expect overlaps between the discriminative concepts that relate to beer aspects.[4] Thus, the five most dissimilar concepts might align less with human-defined concepts.

Table 4: Impact of the number of concepts in ConRAT on the objective performance for the beer reviews.

			Average		
#Concepts		**Acc.**	**P**	**R**	**F**
$\ell=20$	$K=5$	90.95	48.96	37.59	41.37
	$K=10$	**91.35**	**50.02**	**41.96**	**44.86**
	$K=20$	90.24	37.78	31.19	32.84
$\ell=10$	$K=5$	89.64	47.60	19.23	26.90
	$K=10$	**91.25**	**48.12**	**20.11**	**27.97**
	$K=20$	91.05	35.71	14.84	20.71

4.7 RQ 4: How does each module of ConRAT contribute to the overall performance?

Finally, we analyze the importance of each module in an ablation study. To avoid any bias from pruning, we set the number of concepts to five.[5]

Table 5 shows the results. When ConRAT ignores the overlapping or the diversity regularizer, we observe a large drop in the rationale performance. This is expected as the diversity desideratum is not encouraged anymore. However, we remark that the sentiment prediction accuracy increases, which is certainly caused by spurious correlation with the ground-truth label. When all concepts are considered ($s_k = 1 \; \forall k$), we note that the sentiment accuracy stays similar. However, the objective performance decreases by 10% for the precision and more than 20% for the recall and F1 score. These results align with prior work: users write opinions about the topics they care about (Musat and Faltings, 2015; Antognini

[4] As shown in Table 1, the mean length of beer reviews is 184 words. With ℓ=20 and C=20, 400 words are highlighted.

[5] We obtain similar results with K=10 and K=20.

Table 5: Ablation study of ConRAT with five concepts.

		Average		
Model	**Acc.**	**P**	**R**	**F**
ConRAT	89.64	47.60	19.23	26.90
- No $\mathcal{L}_{overlap}$	91.05	31.50	13.16	18.37
- No $\mathcal{L}_{div}$	90.85	34.49	11.69	16.95
- No $s_\theta(\cdot)$:$s_k = 1 \forall k$	89.74	43.13	14.95	21.42
- No Teacher	86.52	45.31	19.65	26.99

et al., 2021a). ConRAT reduces the noise at training by selecting concepts described in the current document. Finally, the teacher model helps ConRAT to boost the sentiment accuracy by more than 3% absolute score, without affecting the rationale quality.

5 Conclusion

Providing explanations for automated predictions carries much more impact, increases transparency, and might even be vital. Previous works have proposed using rationale methods to explain the prediction of a target variable. However, they do not properly capture the multi-faceted nature of useful rationales. We proposed ConRAT, a novel self-explaining model that extracts a set of concepts and explains the outcome with a linear aggregation of concepts, similar to how humans reason.

Our second contribution is two novel regularizers that guide ConRAT to generate interpretable concepts. Experiments on both single- and multi-aspect sentiment classification datasets show that ConRAT, by using only the overall label, is the first to provide superior rationale and predictive performance compared with supervised state-of-the-art methods trained for each aspect label. Moreover, ConRAT produces concepts considered superior in interpretability when evaluated by humans.

References

Stephan Alaniz and Zeynep Akata. 2019. Explainable observer-classifier for explainable binary decisions. *arXiv preprint arXiv:1902.01780*.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7786–7795.

Diego Antognini, Claudiu Musat, and Boi Faltings. 2021a. Interacting with explanations through critiquing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, (IJCAI 2021)*.

Diego Antognini, Claudiu Musat, and Boi Faltings. 2021b. Multi-dimensional explanation of target variables from documents. *Proceedings of the AAAI Conference on Artificial Intelligence, (AAAI 2021)*.

Sharon Lee Armstrong, Lila R. Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition*, 13(3):263 – 308.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Diane Bouchacourt and Ludovic Denoyer. 2019. Educe: Explaining model decisions through unsupervised concepts extraction.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing Systems*, pages 10055–10065.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S Jaakkola. 2020. Invariant rationalization. *arXiv preprint arXiv:2003.09772*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9277–9286.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3543–3556.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*.

Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*, pages 2260–2268.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 1020–1025, Washington, DC, USA.

Claudiu Musat and Boi Faltings. 2015. Personalizing product rankings using collaborative filtering on opinion-derived topic profiles. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Eleanor Quint, Garrett Wirka, Jacob Williams, Stephen Scott, and NV Vinodchandran. 2018. Interpretable classification via supervised variational autoencoders and differentiable decision trees.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153, International Convention Centre, Sydney, Australia. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia. PMLR.

Joshua Brett Tenenbaum. 1999. *A Bayesian framework for concept learning*. Ph.D. thesis, Massachusetts Institute of Technology.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Manzil Zaheer, Amr Ahmed, and Alexander J Smola. 2017. Latent lstm allocation: Joint clustering and non-linear dynamic modeling of sequence data. In *International Conference on Machine Learning*, pages 3967–3976.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Computer Vision – ECCV 2018*, pages 122–138, Cham. Springer International Publishing.

A Additional Training Details

We tune all models on the dev set. We truncate all reviews to 320 tokens for the beer dataset and 400 tokens for Amazon reviews. We have operated a random search over 16 trials. All baselines, except CAR, are tuned for each aspect (80 trials in total for the five aspects). We chose the models achieving the lowest validation accuracy. Most of the time, all models converged under 30 epochs. The range of hyperparameters are the following for ConRAT (similar for other models):

- Learning rate: $[0.0005, 0.00075, 0.001]$;
- Batch size: $[128]$;
- Hidden size: $[256]$;
- λ_D: $[0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1.0]$;
- λ_O: $[0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1.0]$;
- λ_T: $[0.5, 0.6]$;
- Dropout: $[0.0, 0.1]$;
- Weight decay: $[0.0, 10^{-8}, 10^{-10}]$;
- Gumbel temperature in $f_\theta(\cdot)$: $[1.0; 1.5]$;
- Gumbel temperature in $s_\theta(\cdot)$: $[1.0; 1.5]$;

A.1 Hardware / Software

- **CPU**: 2x Intel Xeon E5-2680 v3, 2x 12 cores, 24 threads, 2.5 GHz, 30 MB cache;
- **RAM**: 16x16GB DDR4-2133;
- **GPU**: 2x Nvidia Titan X Maxwell;
- **OS**: Ubuntu 18.04;
- **Software**: Python 3, PyTorch 1.3, CUDA 10.

B Complementary Results RQ 2

B.1 Objective Evaluation

The results for the concept length $\ell = 5$ is shown in Table 6.

Moreover, we report in Table 7 the performance for the **unsupervised** sentiment prediction task for the aspects whose labels are not available to ConRAT: Appearance, Aroma, Palate, and Taste. As we can see, ConRAT achieves competitive results compared to **supervised** baselines.

B.2 Human Evaluation Details

We use Amazon's Mechanical Turk crowdsourcing platform to recruit human annotators to evaluate the quality of extracted justifications and the generated justifications produced by each model. To ensure high-quality of the collected data, we restricted the pool to native English speakers from the U.S., U.K., Canada, or Australia. Additionally, we set the worker requirements at a 98% approval rate and more than 1,000 HITS.

The user interface used to judge the quality of the justifications extracted from different methods, in Section 4.5, is shown in Figure 5.

B.3 Subjective Evaluation

All results (for the joint, the aspect, and the polarity accuracy) are shown in Figure 6. In total, we used 7,500 samples ($100 \times 5 \times 5 \times 3$).

We also studied the error rates on each aspect. The Aroma and Palate aspects cause the highest error for all models. One possible reason is that users confuse these with the aspect Taste, hence their high correlations in rating scores (Antognini et al., 2021b).

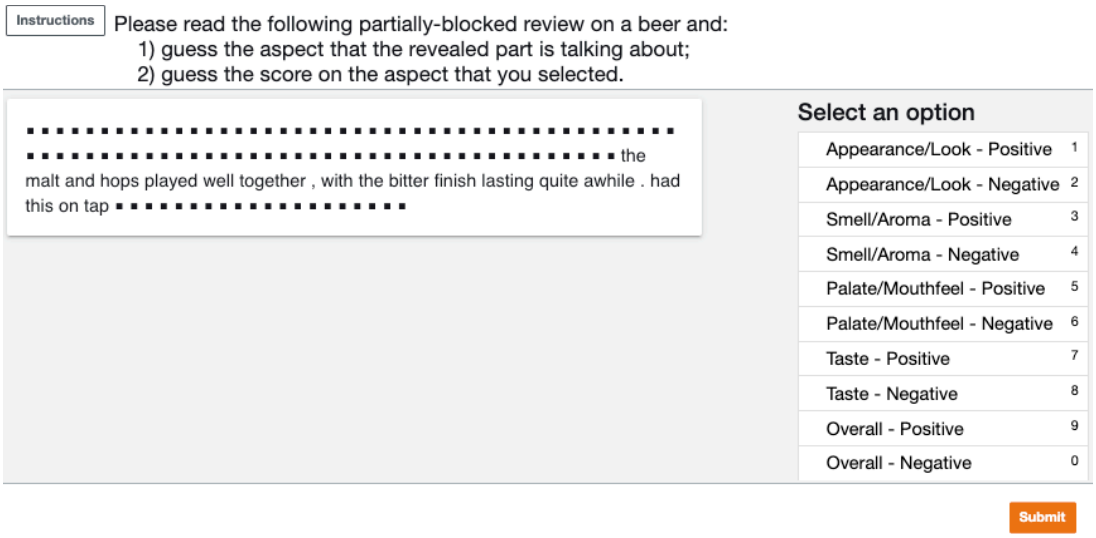
C Extra Visualizations

Additional samples of generated rationales are shown in Figure 7, 8, 9, and 10. We can observe that baselines suffer from spurious correlations: the rationale for the aspect Aroma, Palate, and Taste are often exchanged, or several rationales pick the same text snippets. On the other hand, ConRAT finds better concepts while only trained on the overall aspect label. As it has been shown in prior work (Lei et al., 2016; Chang et al., 2020; Antognini et al., 2021b) rationale methods suffer from the high correlation between rating scores because each model is trained independently for each aspect. Therefore, they rely on the assumption that the data have low internal correlations, which does not reflect the real data distribution. By contrast, ConRAT alleviates this problem be finding all concepts in one training.

Table 6: Objective performance of rationales for the multi-aspect beer reviews with the concept length set to five. ConRAT only uses the overall rating and does not have access to the other aspect labels. All baselines are trained separately on each aspect label. **Bold** and underline denote the best and second-best results, respectively.

			Average			*Appearance*			*Aroma*			*Palate*			*Taste*			*Overall*		
	Model	**Acc.**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
$\ell = 5$	RNP	80.8	41.3	10.4	16.4	50.9	**13.3**	**21.1**	**43.2**	**12.7**	**19.7**	27.1	10.0	14.5	5.5	0.59	1.07	80.0	15.3	25.7
	RNP-3P	81.5	32.9	6.9	11.2	35.1	7.3	12.1	25.6	7.2	11.3	17.0	5.2	8.0	28.6	4.0	7.1	58.2	10.5	17.8
	Intro-3P	84.6	29.8	7.0	11.3	47.3	12.4	19.7	35.4	9.9	15.5	9.7	2.8	4.3	24.3	3.8	6.6	32.4	6.3	10.6
	InvRAT	83.6	46.4	**11.4**	**18.1**	**51.0**	13.1	20.8	40.6	11.9	18.4	**32.0**	**11.8**	**17.2**	36.1	5.6	9.6	72.5	14.7	24.4
	ConRAT†	**90.4**	**46.6**	10.9	17.5	47.2	12.4	19.6	26.9	7.1	11.3	26.6	9.2	13.7	**39.2**	**6.2**	**10.8**	**93.1**	**19.5**	**32.21**

* The model is only trained on the overall label and does not have access to the other ground-truth labels.



Instructions

Please read the following partially-blocked review on a beer and:
1) guess the aspect that the revealed part is talking about;
2) guess the score on the aspect that you selected.

▪▪ the malt and hops played well together , with the bitter finish lasting quite awhile . had this on tap ▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪▪

Select an option

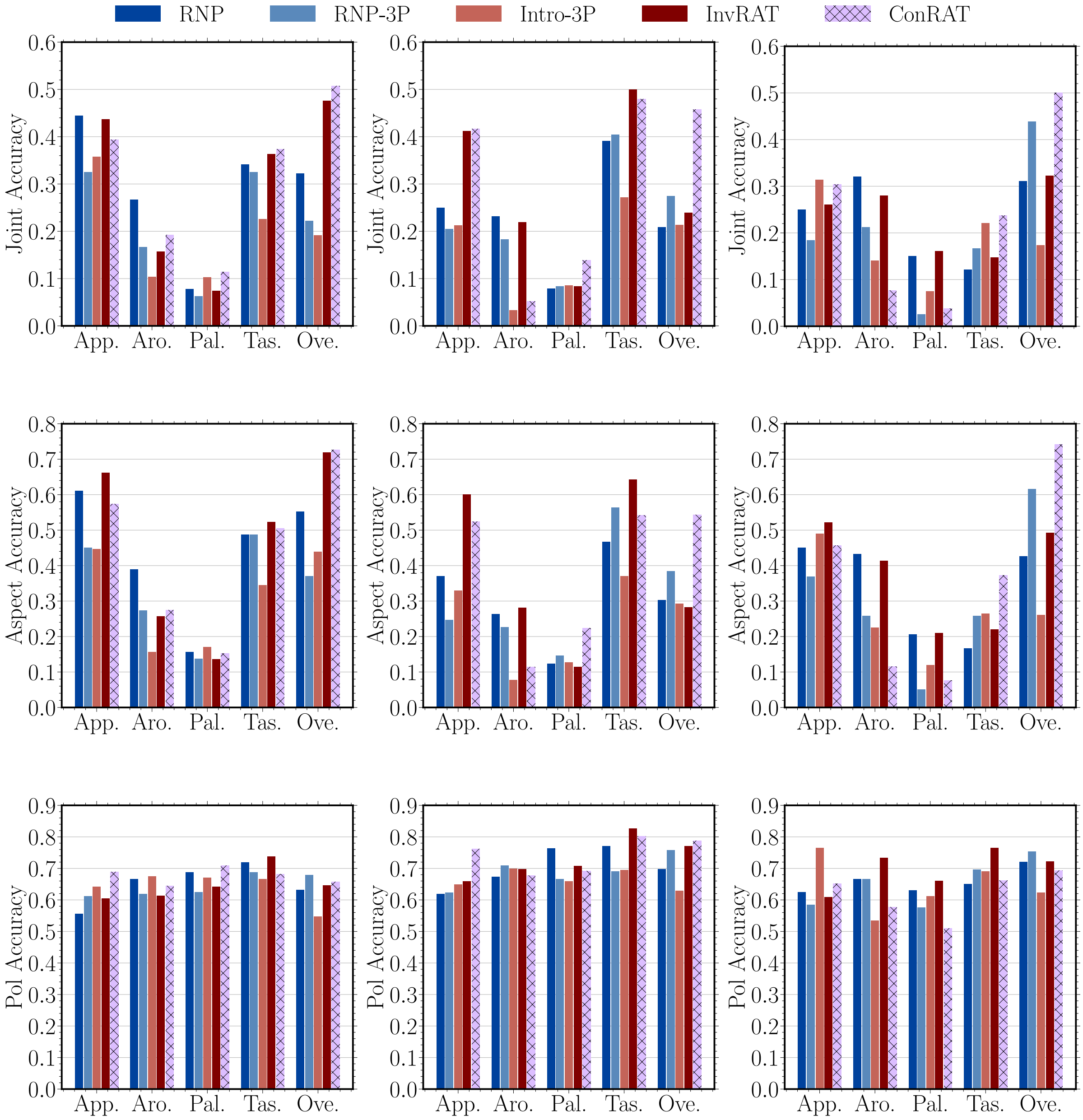
Appearance/Look - Positive	1
Appearance/Look - Negative	2
Smell/Aroma - Positive	3
Smell/Aroma - Negative	4
Palate/Mouthfeel - Positive	5
Palate/Mouthfeel - Negative	6
Taste - Positive	7
Taste - Negative	8
Overall - Positive	9
Overall - Negative	0

Figure 5: Annotation platform for judging the quality of the concepts in the subjective evaluation on beer reviews.

Table 7: Performance on the overall sentiment and the aspects whose labels are not available to ConRAT. **Bold** and underline denote the best and second-best results.

	Model	**Ap.***	**Ar.***	**P***	**T***	**O**
$\ell = 5$	RNP	**95.98**	89.74	92.55	79.78	80.78
	RNP-3P	92.97	87.11	88.09	73.93	81.54
	Intro-3P	93.07	88.38	86.33	77.05	84.57
	InvRAT	**95.98**	90.44	92.66	88.63	83.60
	ConRAT	91.75*	**91.85***	**94.37***	**92.35***	**90.44**
$\ell = 10$	RNP	95.17	**92.15**	**90.74**	82.80	84.41
	RNP-3P	93.55	88.48	90.43	77.15	83.11
	Intro-3P	93.55	87.01	87.21	83.20	80.86
	InvRAT	**95.77**	90.54	89.03	85.01	81.89
	ConRAT	92.25*	91.05*	83.80*	**91.85***	**91.25**
$\ell = 20$	RNP	**96.08**	**92.15**	**94.37**	**87.02**	81.09
	RNP-3P	92.48	87.70	89.16	81.74	80.47
	Intro-3P	93.46	87.11	88.96	86.82	85.64
	InvRAT	95.88	91.05	89.44	85.11	82.90
	ConRAT	67.71*	74.85*	77.16*	80.58*	**91.35**

* ConRAT predicts the sentiment of the aspect in an unsupervised fashion.

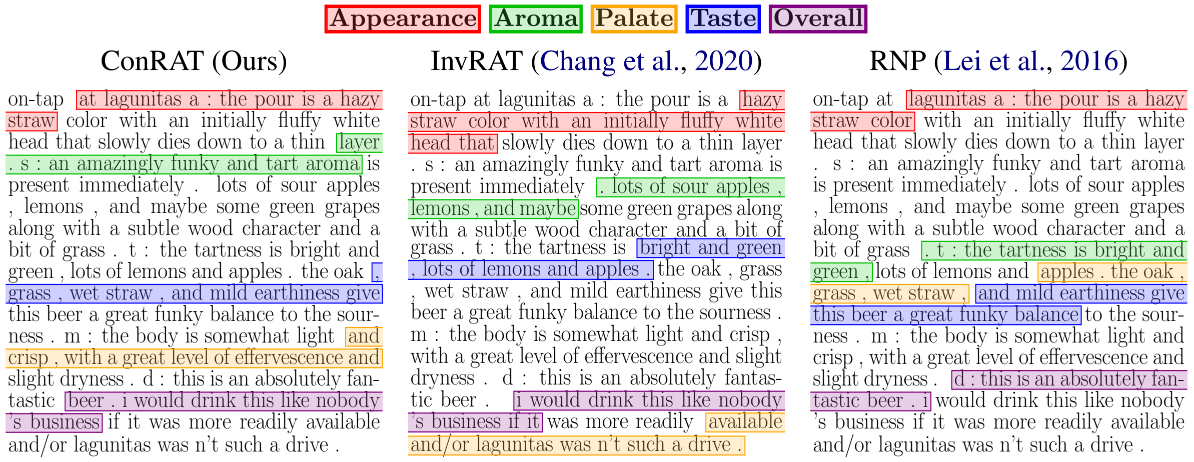


(a) Concept length $\ell = 10$.

(b) Concept length $\ell = 20$.

(c) Concept length $\ell = 5$.

Figure 6: Subjective performance per aspect of rationales for the multi-aspect beer reviews.


Appearance Aroma Palate Taste Overall

ConRAT (Ours)

on-tap at lagunitas a : the pour is a hazy straw color with an initially fluffy white head that slowly dies down to a thin layer . s : an amazingly funky and tart aroma is present immediately . lots of sour apples , lemons , and maybe some green grapes along with a subtle wood character and a bit of grass . t : the tartness is bright and green , lots of lemons and apples . the oak , grass , wet straw , and mild earthiness give this beer a great funky balance to the sourness . m : the body is somewhat light and crisp , with a great level of effervescence and slight dryness . d : this is an absolutely fantastic beer . i would drink this like nobody 's business if it was more readily available and/or lagunitas was n't such a drive .

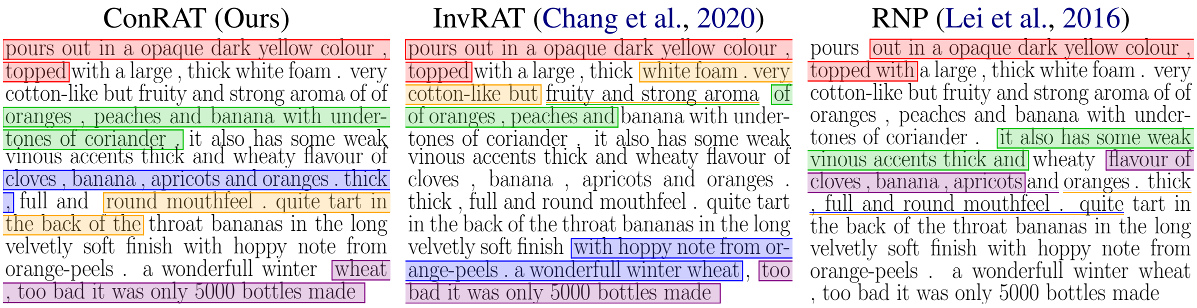
InvRAT (Chang et al., 2020)

on-tap at lagunitas a : the pour is a hazy straw color with an initially fluffy white head that slowly dies down to a thin layer . s : an amazingly funky and tart aroma is present immediately . lots of sour apples , lemons , and maybe some green grapes along with a subtle wood character and a bit of grass . t : the tartness is bright and green , lots of lemons and apples . the oak , grass , wet straw , and mild earthiness give this beer a great funky balance to the sourness . m : the body is somewhat light and crisp , with a great level of effervescence and slight dryness . d : this is an absolutely fantastic beer . i would drink this like nobody 's business if it was more readily available and/or lagunitas was n't such a drive .

RNP (Lei et al., 2016)

on-tap at lagunitas a : the pour is a hazy straw color with an initially fluffy white head that slowly dies down to a thin layer . s : an amazingly funky and tart aroma is present immediately . lots of sour apples , lemons , and maybe some green grapes along with a subtle wood character and a bit of grass . t : the tartness is bright and green , lots of lemons and apples . the oak , grass , wet straw , and mild earthiness give this beer a great funky balance to the sourness . m : the body is somewhat light and crisp , with a great level of effervescence and slight dryness . d : this is an absolutely fantastic beer . i would drink this like nobody 's business if it was more readily available and/or lagunitas was n't such a drive .

Figure 7: Examples of generated rationales with $\ell = 10$ for a beer review. Underline highlights ambiguities.


ConRAT (Ours)

pours out in a opaque dark yellow colour , topped with a large , thick white foam . very cotton-like but fruity and strong aroma of of oranges , peaches and banana with undertones of coriander . it also has some weak vinous accents thick and wheaty flavour of cloves , banana , apricots and oranges . thick , full and round mouthfeel . quite tart in the back of the throat bananas in the long velvetly soft finish with hoppy note from orange-peels . a wonderfull winter wheat , too bad it was only 5000 bottles made

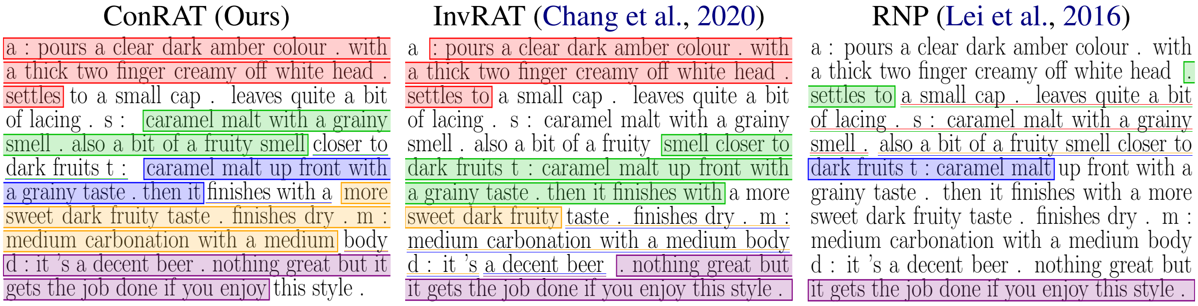
InvRAT (Chang et al., 2020)

pours out in a opaque dark yellow colour , topped with a large , thick white foam . very cotton-like but fruity and strong aroma of of oranges , peaches and banana with undertones of coriander . it also has some weak vinous accents thick and wheaty flavour of cloves , banana , apricots and oranges . thick , full and round mouthfeel . quite tart in the back of the throat bananas in the long velvetly soft finish with hoppy note from orange-peels . a wonderfull winter wheat , too bad it was only 5000 bottles made

RNP (Lei et al., 2016)

pours out in a opaque dark yellow colour , topped with a large , thick white foam . very cotton-like but fruity and strong aroma of of oranges , peaches and banana with undertones of coriander . it also has some weak vinous accents thick and wheaty flavour of cloves , banana , apricots and oranges . thick , full and round mouthfeel . quite tart in the back of the throat bananas in the long velvetly soft finish with hoppy note from orange-peels . a wonderfull winter wheat , too bad it was only 5000 bottles made

Figure 8: Examples of generated rationales with $\ell = 10$ for a beer review. Underline highlights ambiguities.


ConRAT (Ours)

a : pours a clear dark amber colour . with a thick two finger creamy off white head . settles to a small cap . leaves quite a bit of lacing . s : caramel malt with a grainy smell . also a bit of a fruity smell closer to dark fruits t : caramel malt up front with a grainy taste . then it finishes with a more sweet dark fruity taste . finishes dry . m : medium carbonation with a medium body d : it 's a decent beer . nothing great but it gets the job done if you enjoy this style .

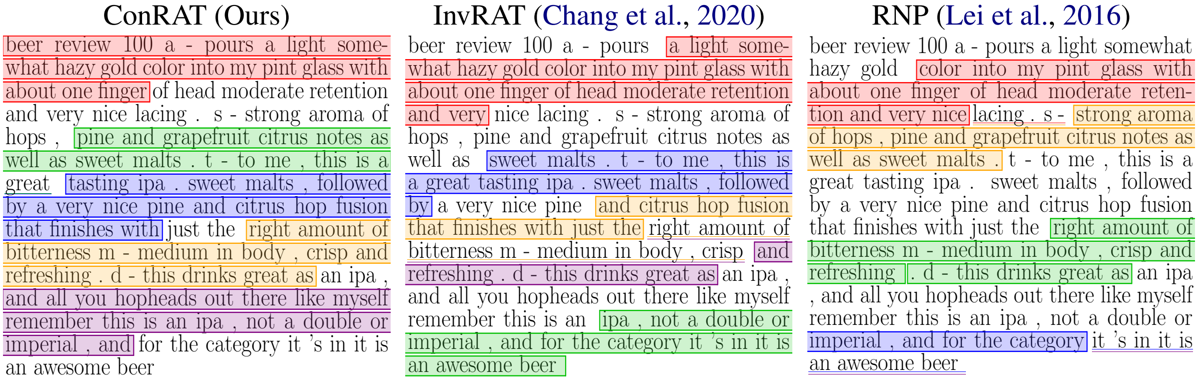
InvRAT (Chang et al., 2020)

a : pours a clear dark amber colour . with a thick two finger creamy off white head . settles to a small cap . leaves quite a bit of lacing . s : caramel malt with a grainy smell . also a bit of a fruity smell closer to dark fruits t : caramel malt up front with a grainy taste . then it finishes with a more sweet dark fruity taste . finishes dry . m : medium carbonation with a medium body d : it 's a decent beer . nothing great but it gets the job done if you enjoy this style .

RNP (Lei et al., 2016)

a : pours a clear dark amber colour . with a thick two finger creamy off white head . settles to a small cap . leaves quite a bit of lacing . s : caramel malt with a grainy smell . also a bit of a fruity smell closer to dark fruits t : caramel malt up front with a grainy taste . then it finishes with a more sweet dark fruity taste . finishes dry . m : medium carbonation with a medium body d : it 's a decent beer . nothing great but it gets the job done if you enjoy this style .

Figure 9: Examples of generated rationales with $\ell = 20$ for a beer review. Underline highlights ambiguities.


ConRAT (Ours)

beer review 100 a - pours a light somewhat hazy gold color into my pint glass with about one finger of head moderate retention and very nice lacing . s - strong aroma of hops , pine and grapefruit citrus notes as well as sweet malts . t - to me , this is a great tasting ipa . sweet malts , followed by a very nice pine and citrus hop fusion that finishes with just the right amount of bitterness m - medium in body , crisp and refreshing . d - this drinks great as an ipa , and all you hopheads out there like myself remember this is an ipa , not a double or imperial , and for the category it 's in it is an awesome beer

InvRAT (Chang et al., 2020)

beer review 100 a - pours a light somewhat hazy gold color into my pint glass with about one finger of head moderate retention and very nice lacing . s - strong aroma of hops , pine and grapefruit citrus notes as well as sweet malts . t - to me , this is a great tasting ipa . sweet malts , followed by a very nice pine and citrus hop fusion that finishes with just the right amount of bitterness m - medium in body , crisp and refreshing . d - this drinks great as an ipa , and all you hopheads out there like myself remember this is an ipa , not a double or imperial , and for the category it 's in it is an awesome beer

RNP (Lei et al., 2016)

beer review 100 a - pours a light somewhat hazy gold color into my pint glass with about one finger of head moderate retention and very nice lacing . s - strong aroma of hops , pine and grapefruit citrus notes as well as sweet malts . t - to me , this is a great tasting ipa . sweet malts , followed by a very nice pine and citrus hop fusion that finishes with just the right amount of bitterness m - medium in body , crisp and refreshing . d - this drinks great as an ipa , and all you hopheads out there like myself remember this is an ipa , not a double or imperial , and for the category it 's in it is an awesome beer

Figure 10: Examples of generated rationales with $\ell = 20$ for a beer review. Underline highlights ambiguities.