Multi-Dimensional Explanation of Target Variables from Documents

Diego Antognini, [1] Claudiu Musat, [2] Boi Faltings, [1]

[1]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[2]Swisscom, Switzerland
firstname.lastname@{epfl.ch, swisscom.com}

Abstract

Automated predictions require explanations to be interpretable by humans. Past work used attention and rationale mechanisms to find words that predict the target variable of a document. Often though, they result in a tradeoff between noisy explanations or a drop in accuracy. Furthermore, rationale methods cannot capture the multi-faceted nature of justifications for multiple targets, because of the non-probabilistic nature of the mask. In this paper, we propose the Multi-Target Masker (MTM) to address these shortcomings. The novelty lies in the soft multi-dimensional mask that models a relevance probability distribution over the set of target variables to handle ambiguities. Additionally, two regularizers guide MTM to induce long, meaningful explanations. We evaluate MTM on two datasets and show, using standard metrics and human annotations, that the resulting masks are more accurate and coherent than those generated by the state-of-the-art methods. Moreover, MTM is the first to also achieve the highest F1 scores for all the target variables simultaneously.

1 Introduction

Neural models have become the standard for natural language processing tasks. Despite the large performance gains achieved by these complex models, they offer little transparency about their inner workings. Thus, their performance comes at the cost of interpretability, limiting their practical utility. Integrating interpretability into a model would supply reasoning for the prediction, increasing its utility.

Perhaps the simplest means of explaining predictions of complex models is by selecting relevant input features. Prior work includes various methods to find relevant words in the text input to predict the target variable of a document. Attention mechanisms (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015) model the word selection by a conditional importance distribution over the inputs, used as explanations to produce a weighted context vector for downstream modules. However, their reliability has been questioned (Jain and Wallace 2019; Pruthi et al. 2020). Another line of research includes rationale generation methods (Lundberg and Lee 2017; Li, Monroe, and Jurafsky 2016; Lei, Barzilay, and Jaakkola 2016). If the selected text input features are short and concise – called a rationale or mask –

Figure 1: A beer review with explanations produced by an attention model and our Multi-Target Masker model. The colors depict produced rationales (i.e., justifications) of the rated aspects: Appearance, Smell, Taste, and Palate. The induced rationales mostly lead to long sequences that clearly describe each aspect (one switch ⋆ per aspect), while the attention model has many short, noisy interleaving sequences.

and suffice on their own to yield the prediction, it can potentially be understood and verified against domain knowledge (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2019). Specifically, these rationale generation methods have been recently proposed to provide such explanations alongside the prediction. Ideally, a good rationale should yield the same or higher performance as using the full input.

The key motivation of our work arises from the limitations of the existing methods. First, the attention mechanisms induce an importance distribution over the inputs, but the resulting explanation consists of many short and noisy word sequences (Figure 1). In addition, the rationale generation methods produce coherent explanations, but the rationales are based on a binary selection of words, leading to the following shortcomings: 1. they explain only one target variable, 2. they make a priori assumptions about the data, and 3. they make it difficult to capture ambiguities in the text. Regarding the first shortcoming, rationales can be multi-faceted by definition

and involve support for different outcomes. If that is the case, one has to train, tune, and maintain one model per target variable, which is impractical. For the second, current models are prone to pick up spurious correlations between the input features and the output. Therefore, one has to ensure that the data have low correlations among the target variables, although this may not reflect the real distribution of the data. Finally, regarding the last shortcoming, a strict assignment of words as rationales might lead to ambiguities that are difficult to capture. For example, in an hotel review that states *"The room was large, clean, and close to the beach."*, the word *"room"* refers to the aspects *Room*, *Cleanliness*, and *Location*. All these limitations are implicitly related due to the non-probabilistic nature of the mask. For further illustrations, see Figure 3 and the appendices.

In this work, we take the best of the attention and rationale methods and propose the Multi-Target Masker to address their limitations by replacing the hard binary mask with a soft multi-dimensional mask (one for each target), in an unsupervised and multi-task learning manner, while jointly predicting all the target variables. We are the first to use a probabilistic multi-dimensional mask to explain multiple target variables jointly without any assumptions on the data, unlike previous rationale generation methods. More specifically, for each word, we model a relevance probability distribution over the set of target variables plus the irrelevant case, because many words can be discarded for every target. Finally, we can control the level of interpretability by two regularizers that guide the model in producing long, meaningful rationales. Compared to existing attention mechanisms, we derive a target importance distribution for each word instead of one over the entire sequence length.

Traditionally, interpretability came at the cost of reduced performance. In contrast, our evaluation shows that on two datasets, in beer and hotel review domains, with up to five correlated targets, our model outperforms strong attention and rationale baselines approaches and generates masks that are strong feature predictors and have a meaningful interpretation. We show that it can be a benefit to: 1. guide the model to focus on different parts of the input text, 2. capture ambiguities of words belonging to multiple aspects, and 3. further improve the sentiment prediction for all the aspects. Thus, interpretability does not come at a cost in our paradigm.

2 Related Work

2.1 Interpretability

Developing interpretable models is of considerable interest to the broader research community; this is even more pronounced with neural models (Kim, Shah, and Doshi-Velez 2015; Doshi-Velez and Kim 2017). There has been much work with a multitude of approaches in the areas of analyzing and visualizing state activation (Karpathy, Johnson, and Li 2015; Li et al. 2016; Montavon, Samek, and Müller 2018), attention weights (Jain and Wallace 2019; Serrano and Smith 2019; Pruthi et al. 2020), and learned sparse and interpretable word vectors (Faruqui et al. 2015b,a; Herbelot and Vecchi 2015). Other works interpret black box models by locally fitting interpretable models (Ribeiro, Singh, and Guestrin

2016; Lundberg and Lee 2017). (Li, Monroe, and Jurafsky 2016) proposed erasing various parts of the input text using reinforcement learning to interpret the decisions. However, this line of research aims at providing post-hoc explanations of an already-trained model. Our work differs from these approaches in terms of what is meant by an explanation and its computation. We defined an explanation as one or multiple text snippets that – as a substitute of the input text – are sufficient for the predictions.

2.2 Attention-based Models

Attention models (Vaswani et al. 2017; Yang et al. 2016; Lin et al. 2017) have been shown to improve prediction accuracy, visualization, and interpretability. The most popular and widely used attention mechanism is soft attention (Bahdanau, Cho, and Bengio 2015), rather than hard attention (Luong, Pham, and Manning 2015) or sparse ones (Martins and Astudillo 2016). According to various studies (Jain and Wallace 2019; Serrano and Smith 2019; Pruthi et al. 2020), standard attention modules noisily predict input importance; the weights cannot provide safe and meaningful explanations. Moreover, (Pruthi et al. 2020) showed that standard attention modules can fool people into thinking that predictions from a model biased against gender minorities do not rely on the gender. Our approach differs in two ways from attention mechanisms. First, the loss includes two regularizers to favor long word sequences for interpretability. Second, the normalization is not done over the sequence length but over the target set for each word; each has a relevance probability distribution over the set of target variables.

2.3 Rationale Models

The idea of including human rationales during training has been explored in (Zhang, Marshall, and Wallace 2016; Bao et al. 2018; DeYoung et al. 2020). Although they have been shown to be beneficial, they are costly to collect and might vary across annotators. In our work, no annotation is needed.

One of the first rationale generation methods was introduced by (Lei, Barzilay, and Jaakkola 2016) in which a generator masks the input text fed to the classifier. This framework is a cooperative game that selects rationales to accurately predict the label by maximizing the mutual information (Chen et al. 2018). (Yu et al. 2019) proposed conditioning the generator based on the predicted label from a classifier reading the whole input, although it slightly underperformed when compared to the original model (Chang et al. 2020). (Chang et al. 2019) presented a variant that generated rationales to perform counterfactual reasoning. Finally, (Chang et al. 2020) proposed a generator that can decrease spurious correlations in which the selective rationale consists of an extracted chunk of a pre-specified length, an easier variant than the original one that generated the rationale. In all, these models are trained to generate a hard binary mask as a rationale to explain the prediction of a target variable, and the method requires as many models to train as variables to explain. Moreover, they rely on the assumption that the data have low internal correlations.

In contrast, our model addresses these drawbacks by jointly predicting the rationales of all the target variables (even in
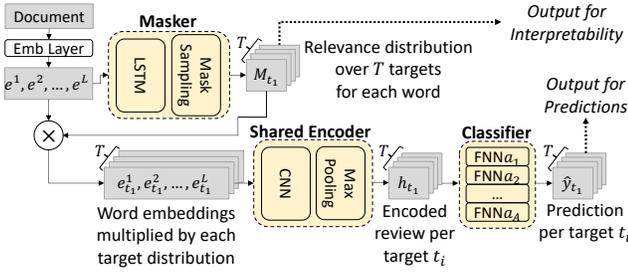
Figure 2: The proposed Multi-Target Masker (*MTM*) model architecture to predict and explain T target variables.

the case of highly correlated data) by generating a soft multi-dimensional mask. The probabilistic nature of the masks can handle ambiguities in the induced rationales. In our recent work (Antognini, Musat, and Faltings 2020), we show how to use the induced rationales to generate personalized explanations for recommendation and how human users significantly prefer these over those produced by state-of-the-art models.

3 The Multi-Target Masker (MTM)

Let X be a random variable representing a document composed of L words $(x^1, x^2, ..., x^L)$, and Y the target T-dimensional vector.[1] Our proposed model, called the Multi-Target Masker (MTM), is composed of three components: 1) a **masker** module that computes a probability distribution over the target set for each word, resulting in $T+1$ masks (including one for the irrelevant case); 2) an **encoder** that learns a representation of a document X conditioned on the induced masks; 3) a **classifier** that predicts the target variables. The overall model architecture is shown in Figure 2. Each module is interchangeable with other models.

3.1 Model Overview

Masker. The masker first computes a hidden representation h^ℓ for each word x^ℓ in the input sequence, using their word embeddings $e^1, e^2, ..., e^L$. Many sequence models could realize this task, such as recurrent, attention, or convolution networks. In our case, we chose a recurrent model to learn the dependencies between the words. Let t_i be the ith target for $i = 1, ..., T$, and t_0 the irrelevant case, because many words are irrelevant to every target. We define the multi-dimensional mask $\mathbf{M} \in \mathbb{R}^{(T+1) \times L}$ as the target relevance distribution $M^\ell \in \mathbb{R}^{(T+1)}$ of each word x^ℓ as follows:

$$P(\mathbf{M}|X) = \prod_{\ell=1}^{L} P(M^\ell|x^\ell) = \prod_{\ell=1}^{L} \prod_{i=0}^{T} P(m_{t_i}^\ell|x^\ell) \quad (1)$$

Because we have categorical distributions, we cannot directly sample $P(M^\ell|x^\ell)$ and backpropagate the gradient through this discrete generation process. Instead, we model the variable $m_{t_i}^\ell$ using the straight through gumbel-softmax (Jang, Gu, and Poole 2017; Maddison, Mnih, and

Teh 2017) to approximate sampling from a categorical distribution.[2] We model the parameters of each Gumbel-Softmax distribution M^ℓ with a single-layer feed-forward neural network followed by applying a log softmax, which induces the log-probabilities of the ℓth distribution: $\omega_\ell = \log(\text{softmax}(Wh^\ell + b))$. W and b are shared across all tokens so that the number of parameters stays constant with respect to the sequence length. We control the sharpness of the distributions with the temperature parameter τ, which dictates the peakiness of the relevance distributions. In our case, we keep the temperature low to enforce the assumption that each word is relevant about one or two targets. Note that compared to attention mechanisms, the word importance is a probability distribution over the targets $\sum_{i=0}^{T} P(m_{t_i}^\ell|x^\ell) = 1$ instead of a normalization over the sequence length $\sum_{\ell=1}^{L} P(t^\ell|x^\ell) = 1$.

Given a soft multi-dimensional mask $\mathbf{M} \in \mathbb{R}^{(T+1) \times L}$, we define each sub-mask $M_{t_i} \in \mathbb{R}^L$ as follows:

$$M_{t_i} = P(m_{t_i}^1|x^1), P(m_{t_i}^2|x^2), ..., P(m_{t_i}^L|x^L) \quad (2)$$

To integrate the word importance of the induced sub-masks M_{t_i} within the model, we weight the word embeddings by their importance towards a target variable t_i, such that $E_{t_i} = E \odot M_{t_i} = e_1 \cdot P(m_{t_i}^1|x^1), e_2 \cdot P(m_{t_i}^2|x^2), ..., e_L \cdot P(m_{t_i}^L|x^L)$. Thereafter, each modified embedding E_{t_i} is fed into the encoder block. Note that E_{t_0} is ignored because M_{t_0} only serves to absorb probabilities of words that are insignificant.[3]

Encoder and Classifier. The encoder includes a convolutional network, followed by max-over-time pooling to obtain a fixed-length feature vector. We chose a convolutional network because it led to a smaller model, faster training, and performed empirically similarly to recurrent and attention models. It produces the fixed-size hidden representation h_{t_i} for each target t_i. To exploit commonalities and differences among the targets, we share the weights of the encoder for all E_{t_i}. Finally, the classifier block contains for each target variable t_i a two-layer feedforward neural network, followed by a softmax layer to predict the outcome $\hat{y}_{t_i}$.

Extracting Rationales. To explain the prediction $\hat{y}_{t_i}$ of one target Y_{t_i}, we generate its rationale by selecting each word x^ℓ, whose relevance towards t_i is the most likely: $P(m_{t_i}^\ell|x^\ell) = \max_{j=0,...,T} P(m_{t_j}^\ell|x^\ell)$. Then, we can interpret $P(m_{t_i}^\ell|x^\ell)$ as the model confidence of x^ℓ relevant to Y_{t_i}.

3.2 Enabling the Interpretability of Masks

The first objective to optimize is the prediction loss, represented as the cross-entropy between the true target label y_{t_i} and the prediction $\hat{y}_{t_i}$ as follows:

$$\ell_{pred} = \sum_{i=1}^{T} \ell_{cross\_entropy}(y_{t_i}, \hat{y}_{t_i}) \quad (3)$$

---

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

[2]We also experimented with the implicit reparameterization trick using a Dirichlet distribution (Figurnov, Mohamed, and Mnih 2018) instead, but we did not obtain a significant improvement.

[3]if $P(m_{t_0}^\ell|x^\ell) \approx 1.0$, it implies $\sum_{i=1}^{T} P(m_{t_i}^\ell|x^\ell) \approx 0$ and consequently, $e_{t_i}^\ell \approx \vec{0}$ for $i = 0, ..., T$.

However, training MTM to optimize ℓ_{pred} will lead to meaningless sub-masks M_{t_i} because the model tends to focus on certain words. Consequently, we guide the model to produce long, meaningful word sequences, as shown in Figure 1. We propose two regularizers to control the number of selected words and encourage consecutive words to be relevant to the same target. For the first term, we calculate the probability p_{sel} of tagging a word as relevant to any target as follows:

$$p_{sel} = \frac{1}{L} \sum_{\ell=1}^{L} \left(1 - P(m_{t_0}^{\ell}|x^{\ell})\right) \qquad (4)$$

We then compute the cross-entropy with a prior hyperparameter λ_p to control the expected number of selected words among all target variables, which corresponds to the expectation of a binomial distribution (p_{sel}). We minimize the difference between p_{sel} and λ_p as follows:

$$\ell_{sel} = \ell_{binary\_cross\_entropy}(p_{sel}, \lambda_p) \qquad (5)$$

The second regularizer discourages the target transition of two consecutive words by minimizing the mean variation of their target distributions, M^{ℓ} and $M^{\ell-1}$. We generalize the formulation of a hard binary selection as suggested by (Lei, Barzilay, and Jaakkola 2016) to a soft probabilistic multi-target selection as follows:[4]

$$p_{dis} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{||M^{\ell} - M^{\ell-1}||_1}{A + 1} \qquad (6)$$

$$\ell_{cont} = \ell_{binary\_cross\_entropy}(p_{dis}, 0)$$

We train our Multi-Target Masker end to end and optimize the loss $\ell_{MTM} = \ell_{pred} + \lambda_{sel} \cdot \ell_{sel} + \lambda_{cont} \cdot \ell_{cont}$, where λ_{sel} and λ_{cont} control the impact of each constraint.

4 Experiments

We assess our model in two dimensions: the quality of the explanations, obtained from the masks, and the predictive performance. Following previous work (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2020), we use sentiment analysis as a demonstration use case, but we extend it to the multi-aspect case. However, we are interested in learning rationales for every aspect at the same time without any prior assumption on the data, where aspect ratings can be highly correlated. We first measure the quality of the induced rationales using human aspect sentence-level annotations and an automatic topic model evaluation method. In the second set of experiments, we evaluate MTM on the multi-aspect sentiment classification task in two different domains.[5]

4.1 Experimental Details

The review encoder was either a bi-directional recurrent neural network using LSTM (Hochreiter and Schmidhuber 1997) with 50 hidden units or a multi-channel text convolutional neural network, similar to (Kim, Shah, and Doshi-Velez

---

[4]Early experiments with other distance functions, such as the Kullback–Leibler divergence, produced inferior results.

[5]Code & data available at https://github.com/Diego999/MTM.

Dataset	Beer	Hotel
Number of reviews	$1,586,259$	$140,000$
Average words per review	147.1 ± 79.7	188.3 ± 50.0
Average sentences per review	10.3 ± 5.4	10.4 ± 4.4
Number of Aspects	4	5
Avg./Max. corr. between aspects	$71.8\%/73.4\%$	$63.0\%/86.5\%$

Table 1: Statistics of the multi-aspect review datasets. Both datasets have high correlations between aspects.

2015), with 3-, 5-, and 7-width filters and 50 feature maps per filter. Each aspect classifier is a two-layer feedforward neural network with a rectified linear unit activation function (Nair and Hinton 2010). We used the 200-dimensional pre-trained word embeddings of (Lei, Barzilay, and Jaakkola 2016) for beer reviews. For the hotel domain, we trained word2vec (Mikolov et al. 2013) on a large collection of hotel reviews (Antognini and Faltings 2020) with an embedding size of 300. We used a dropout (Srivastava et al. 2014) of 0.1, clipped the gradient norm at 1.0, added a L2-norm regularizer with a factor of 10^{-6}, and trained using early stopping. We used Adam (Kingma and Ba 2015) with a learning rate of 0.001. The temperature τ for the Gumbel-Softmax distributions was fixed at 0.8. The two regularizers and the prior of our model were $\lambda_{sel} = 0.03$, $\lambda_{cont} = 0.03$, and $\lambda_p = 0.15$ for the *Beer* dataset and $\lambda_{sel} = 0.02$, $\lambda_{cont} = 0.02$, and $\lambda_p = 0.10$ for the *Hotel* one. We ran all experiments for a maximum of 50 epochs with a batch-size of 256. We tuned all models on the dev set with 10 random search trials.

4.2 Datasets

(McAuley, Leskovec, and Jurafsky 2012) provided 1.5 million English beer reviews from BeerAdvocat. Each contains multiple sentences describing various beer aspects: *Appearance*, *Smell*, *Palate*, and *Taste*; users also provided a five-star rating for each aspect. To evaluate the robustness of the models across domains, we sampled $140\,000$ hotel reviews from (Antognini and Faltings 2020), that contains 50 million reviews from TripAdvisor. Each review contains a five-star rating for each aspect: *Service*, *Cleanliness*, *Value*, *Location*, and *Room*. The descriptive statistics are shown in Table 1.

There are high correlations among the rating scores of different aspects in the same review (71.8% and 63.0% on average for the beer and hotel datasets, respectively). This makes it difficult to directly learn textual justifications for single-target rationale generation models (Chang et al. 2020, 2019; Lei, Barzilay, and Jaakkola 2016). Prior work used separate decorrelated train sets for each aspect and excluded aspects with a high correlation, such as *Taste*, *Room*, and *Value*. However, these assumptions do not reflect the real data distribution. Therefore, we keep the original data (and thus can show that our model does not suffer from the high correlations). We binarize the problem as in previous work (Bao et al. 2018; Chang et al. 2020): ratings at three and above are labeled as positive and the rest as negative. We split the data into 80/10/10 for the train, validation, and test sets. Compared to the beer reviews, the hotel ones were longer, noisier, and less structured, as shown in Appendices A.3 and A.2.

4.3 Baselines

We compare our Multi-Target Masker (*MTM*) with various baselines. We group them in three levels of interpretability:

- *None*. We cannot extract the input features the model used to make the predictions;

- *Coarse-grained*. We can observe what parts of the input a model used to discriminate all aspect sentiments without knowing what part corresponded to what aspect;

- *Fine-grained*. For each aspect, a model selects input features to make the prediction.

We first use a simple baseline, *SENT*, that reports the majority sentiment across the aspects, as the aspect ratings are highly correlated. Because this information is not available at testing, we trained a model to predict the majority sentiment of a review as suggested by (Wang and Manning 2012). The second baseline we used is a shared encoder followed by T classifiers that we denote *BASE*. These models do not offer any interpretability. We extend it with a shared attention mechanism (Bahdanau, Cho, and Bengio 2015) after the encoder, noted as *SAA* in our study, that provides a coarse-grained interpretability; for all aspects, *SAA* focuses on the same words in the input.

Our final goal is to achieve the best performance and provide fine-grained interpretability in order to visualize what sequences of words a model focuses on to predict the aspect sentiments. To this end, we include other baselines: two trained *separately* for each aspect (e.g., current rationale models) and two trained with a *multi-aspect* sentiment loss. For the first ones, we employ the the well-known *NB-SVM* (Wang and Manning 2012) for sentiment analysis tasks, and we then use the Single-Aspect Masker (*SAM*) (Lei, Barzilay, and Jaakkola 2016), each trained separately for each aspect.

The two last methods contain a separate encoder, attention mechanism, and classifier for each aspect. We utilize two types of attention mechanisms, additive (Bahdanau, Cho, and Bengio 2015) and sparse (Martins and Astudillo 2016), as sparsity in the attention has been shown to induce useful, interpretable representations. We call them Multi-Aspect Attentions (*MAA*) and Sparse-Attentions (*MASA*), respectively. Diagrams of the baselines can be found in Appendix A.4.

We demonstrate that the induced sub-masks $M_{t_i}, ..., M_{t_T}$ computed from *MTM*, bring fine-grained interpretability and are meaningful for other models to improve performance. To do so, we extract and concatenate the masks to the word embeddings, resulting in contextualized embeddings (Peters et al. 2018), and train *BASE* with those. We call this variant MTM^C, that is smaller and has faster inference than *MTM*.

5 Results

5.1 Multi-Rationale Interpretability

We first verify whether the inferred rationales of *MTM* are meaningful and interpretable, compared to the other models.

Precision. Evaluating explanations that consist of coherent pieces of text is challenging because there is no gold standard for reviews. (McAuley, Leskovec, and Jurafsky 2012) have provided 994 beer reviews with sentence-level aspect

	Precision / % Highlighted Words		
Model	*Smell*	*Palate*	*Appearance*
NB-SVM[*]	21.6 / 7%	24.9 / 7%	38.3 / 13%
SAA[*]	88.4 / 7%	65.3 / 7%	80.6 / 13%
SAM[*]	95.1 / 7%	80.2 / 7%	96.3 / 14%
MASA	87.0 / 4%	42.8 / 5%	74.5 / 4%
MAA	51.3 / 7%	32.9 / 7%	44.9 / 14%
MTM	**96.6** / 7%	**81.7** / 7%	**96.7** / 14%

[*] Model trained separately for each aspect.

Table 2: Performance related to human evaluation, showing the precision of the selected words for each aspect of the *Beer* dataset. The percentage of words indicates the number of highlighted words of the full review.

annotations (although our model computes masks at a finer level). Each sentence was annotated with one aspect label, indicating what aspect that sentence covered. We evaluate the precision of the words selected by each model, as in (Lei, Barzilay, and Jaakkola 2016). We use trained models on the *Beer* dataset and extracted a similar number of selected words for a fair comparison. We also report the results of the models from (Lei, Barzilay, and Jaakkola 2016): *NB-SVM*, the Single-Aspect Attention and Masker (*SAA* and *SAM*, respectively); they use the separate decorrelated train sets for each aspect because they compute hard masks.[6]

Table 2 presents the precision of the masks and attentions computed on the sentence-level aspect annotations. We show that the generated sub-masks obtained with our Multi-Target Masker (*MTM*) correlates best with the human judgment. In comparison to *SAM*, the *MTM* model obtains significantly higher precision with an average of $+1.13$. Interestingly, *NB-SVM* and attention models (*SAA*, *MASA*, and *MAA*) perform poorly compared with the mask models, especially *MASA*, which focuses only on a couple of words due to the sparseness of the attention. In Appendix D, we also analyze the impact of the length of the explanations.

Semantic Coherence. In addition to evaluating the rationales with human annotations, we compute their semantic interpretability. According to (Aletras and Stevenson 2013; Lau, Newman, and Baldwin 2014), normalized point mutual information (NPMI) is a good metric for the qualitative evaluation of topics because it matches human judgment most closely. However, the top-N topic words used for evaluation are often selected arbitrarily. To alleviate this problem, we followed (Lau and Baldwin 2016). We compute the topic coherence over several cardinalities and report the results and average (see Appendix A.1); those authors claimed that the mean leads to a more stable and robust evaluation.

The results are shown in Table 3. We show that the computed masks by *MTM* lead to the highest mean NPMI and, on average, 20% superior results in both datasets, while only needing a single training. Our *MTM* model significantly outperforms *SAM* and the attention models (*MASA* and *MAA*) for $N \geq 20$ and $N = 5$. For $N = 10$ and $N = 15$, *MTM*

---

[6]When trained on the original data, they performed significantly worse, showing the limitation in handling correlated variables.

	NPMI						
Model	$N=5$	10	15	20	25	30	Mean[†]
			Beer				
SAM[*]	0.046	0.120	0.129	0.243	0.308	0.396	0.207
MASA	0.020	0.082	0.130	0.168	0.234	0.263	0.150
MAA	0.064	**0.189**	0.255	0.273	0.332	0.401	0.252
MTM	**0.083**	0.187	**0.264**	**0.348**	**0.477**	**0.410**	**0.295**
			Hotel				
SAM[*]	0.041	0.103	0.152	0.180	0.233	0.281	0.165
MASA	0.043	0.127	0.166	0.295	0.323	0.458	0.235
MAA	0.128	0.218	**0.352**	0.415	0.494	0.553	0.360
MTM	**0.134**	**0.251**	0.349	**0.496**	**0.641**	**0.724**	**0.432**

[*] Model trained separately for each aspect.

[†] The metric that correlates best with human judgment (Lau and Baldwin 2016).

Table 3: Performance on automatic evaluation, showing the average topic coherence (NPMI) across different top-N words for each dataset. We considered each aspect a_i as a topic and used the masks/attentions to compute $P(w|a_i)$.

obtains higher scores in two out of four cases ($+.033$ and $+.009$). For the other two, the difference was below $.003$. SAM obtains poor results in all cases.

We analyzed the top words for each aspect by conducting a human evaluation to identify intruder words (i.e., words not matching the corresponding aspect). Generally, our model found better topic words: approximately 1.9 times fewer intruders than other methods for each aspect and each dataset. More details are available in Appendix A.1.

5.2 Multi-Aspect Sentiment Classification

We showed that the inferred rationales of *MTM* were significantly more accurate and semantically coherent than those produced by the other models. Now, we inquire as to whether the masks could become a benefit rather than a cost in performance for the multi-aspect sentiment classification.

Beer Reviews. We report the macro F1 and individual score for each aspect A_i. Table 4 (top) presents the results for the *Beer* dataset. The Multi-Target Masker (*MTM*) performs better on average than all the baselines and provided fine-grained interpretability. Moreover, *MTM* has two times fewer parameters than the aspect-wise attention models.

The contextualized variant MTM^C achieves a macro F1 score absolute improvement of 0.44 and 2.49 compared to *MTM* and *BASE*, respectively. These results highlight that the inferred masks are meaningful to improve the performance while bringing fine-grained interpretability to *BASE*. It is 1.5 times smaller than *MTM* and has a faster inference.

NB-SVM, which offers fine-grained interpretability and was trained separately for each aspect, significantly underperforms when compared to *BASE* and, surprisingly, to *SENT*. As shown in Table 1, the sentiment correlation between any pair of aspects of the *Beer* dataset is on average 71.8%. Therefore, by predicting the sentiment of one aspect correctly, it is likely that other aspects share the same polarity. We suspect

Service Cleanliness Value Location Room

Multi-Target Masker (Ours)

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean , good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! [...] plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro ! would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Single-Aspect Masker

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! [...] plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Multi-Aspect Attentions

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! [...] plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Multi-Aspect Sparse-Attentions

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! [...] plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Figure 3: Induced rationales on a truncated hotel review, where shade colors represent the model confidence towards the aspects. *MTM* finds most of the crucial spans of words with a small amount of noise. *SAM* lacks coverage but identifies words where half are correct and the others ambiguous (represented with colored underlines).

that the linear model *NB-SVM* cannot capture the correlated relationships between aspects, unlike the non-linear (neural) models that have a higher capacity. The shared attention models perform better than *BASE* but provide only coarse-grained interpretability. *SAM* is outperformed by all the models except *SENT*, *BASE*, and *NB-SVM*.

Model Robustness - Hotel Reviews. We check the robustness of our model on another domain. Table 4 (bottom) presents the results of the *Hotel* dataset. The contextualized variant MTM^C outperforms all other models significantly with a macro F1 score improvement of 0.49. Moreover, it achieves the best individual F1 score for each aspect A_i. This shows that the learned mask $\mathbf{M}$ of *MTM* is again meaningful because it increases the performance and adds interpretability to *BASE*. Regarding *MTM*, we see that it performs slightly worse than the aspect-wise attention models *MASA* and *MAA* but has 2.5 times fewer parameters.

A visualization of a truncated hotel review with the extracted rationales and attentions is available in Figure 3. Not only do probabilistic masks enable higher performance, they better capture parts of reviews related to each aspect compared to other methods. More samples of beer and hotel reviews can be found in Appendix A.3.

To summarize, we have shown that the regularizers in

				F1 Scores				
Interp.		**Model**	**Params**	**Macro**	A_1	A_2	A_3	A_4
None	SENT	Sentiment Majority	$560k$	73.01	71.83	75.65	71.26	73.31
	BASE	$Emb_{200} + Enc_{CNN} + Clf$	$188k$	76.45	71.44	78.64	74.88	80.83
Coarse-grained	SAA	$Emb_{200} + Enc_{CNN} + A_{Shared} + Clf$	$226k$	77.06	73.44	78.68	75.79	80.32
		$Emb_{200} + Enc_{LSTM} + A_{Shared} + Clf$	$219k$	78.03	74.25	79.53	75.76	82.57
Fine-grained	NB-SVM	(Wang and Manning 2012)	$4 \cdot 560k$	72.11	72.03	74.95	68.11	73.35
	SAM	(Lei, Barzilay, and Jaakkola 2016)	$4 \cdot 644k$	76.62	72.93	77.94	75.70	79.91
	MASA	$Emb_{200} + Enc_{LSTM} + A_{Aspect-wise}^{Sparse} + Clf$	$611k$	77.62	72.75	79.62	75.81	82.28
	MAA	$Emb_{200} + Enc_{LSTM} + A_{Aspect-wise} + Clf$	$611k$	78.50	74.58	79.84	77.06	82.53
	MTM	$Emb_{200} + Masker + Enc_{CNN} + Clf$ (Ours)	$289k$	78.55	74.87	79.93	77.39	82.02
	MTM^C	**$Emb_{200+4} + Enc_{CNN} + Clf$ (Ours)**	$191k$	**78.94**	**75.02**	**80.17**	**77.86**	**82.71**

Beer Reviews (left label for top section)

				F1 Scores					
Interp.		**Model**	**Params**	**Macro**	A_1	A_2	A_3	A_4	A_5
None	SENT	Sentiment Majority	$309k$	85.91	89.98	90.70	92.12	65.09	91.67
	BASE	$Emb_{300} + Enc_{CNN} + Clf$	$263k$	90.30	92.91	93.55	94.12	76.65	94.29
Coarse-grained	SAA	$Emb_{300} + Enc_{CNN} + A_{Shared} + Clf$	$301k$	90.12	92.73	93.55	93.76	76.40	94.17
		$Emb_{300} + Enc_{LSTM} + A_{Shared} + Clf$	$270k$	88.22	91.13	92.19	93.33	71.40	93.06
Fine-grained	NB-SVM	(Wang and Manning 2012)	$5 \cdot 309k$	87.17	90.04	90.77	92.30	71.27	91.46
	SAM	(Lei, Barzilay, and Jaakkola 2016)	$5 \cdot 824k$	87.52	91.48	91.45	92.04	70.80	91.85
	MASA	$Emb_{200} + Enc_{LSTM} + A_{Aspect-wise}^{Sparse} + Clf$	$1010k$	90.23	93.11	93.32	93.58	77.21	93.92
	MAA	$Emb_{300} + Enc_{LSTM} + A_{Aspect-wise} + Clf$	$1010k$	90.21	92.84	93.34	93.78	76.87	94.21
	MTM	$Emb_{300} + Masker + Enc_{CNN} + Clf$ (Ours)	$404k$	89.94	92.84	92.95	93.91	76.27	93.71
	MTM^C	**$Emb_{300+5} + Enc_{CNN} + Clf$ (Ours)**	$267k$	**90.79**	**93.38**	**93.82**	**94.55**	**77.47**	**94.71**

Hotel Reviews (left label for bottom section)

Table 4: Performance of the multi-aspect sentiment classification task for the *Beer* (top) and *Hotel* (bottom) datasets.

MTM guide the model to produce high-quality masks as explanations while performing slightly better than the strong attention models in terms of prediction performance. However, we demonstrated that including the inferred masks into word embeddings and training a simpler model achieved the best performance across two datasets and and at the same time, brought fine-grained interpretability. Finally, *MTM* supported high correlation among multiple target variables.

Hard Mask versus Soft Masks. *SAM* is the neural model that obtained the lowest relative macro F1 score in the two datasets compared with MTM^C: a difference of -2.32 and -3.27 for the *Beer* and *Hotel* datasets, respectively. Both datasets have a high average correlation between the aspect ratings: 71.8% and 63.0%, respectively (see Table 1). Therefore, it makes it challenging for rationale models to learn the justifications of the aspect ratings directly. Following the observations of (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2019, 2020), this highlights that single-target rationale models suffer from high correlations and require data to satisfy certain constraints, such as low correlations. In contrast, *MTM* does not require any particular assumption on the data.

We compare *MTM* in a setting where the aspect ratings were less correlated, although it does not reflect the real distribution of the aspect ratings. We employ the decorrelated subsets of the *Beer* reviews from (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2020). It has an average correlation of 27.2% and the aspect *Taste* is removed.

We find similar trends but stronger results: *MTM* significantly generates better rationales and achieves higher F1 scores than *SAM* and the attention models. The contextualized variant MTM^C further improves the performance. The full results and visualizations are available in Appendix A.2.

6 Conclusion

Providing explanations for automated predictions carries much more impact, increases transparency, and might even be necessary. Past work has proposed using attention mechanisms or rationale methods to explain the prediction of a target variable. The former produce noisy explanations, while the latter do not properly capture the multi-faceted nature of useful rationales. Because of the non-probabilistic assignment of words as justifications, rationale methods are prone to suffer from ambiguities and spurious correlations and thus, rely on unrealistic assumptions about the data.

The Multi-Target Masker (MTM) addresses these drawbacks by replacing the binary mask with a probabilistic multi-dimensional mask (one dimension per target), learned in an unsupervised and multi-task learning manner, while jointly predicting all the target variables.

According to comparison with human annotations and automatic evaluation on two real-world datasets, the inferred masks were more accurate and coherent than those that were produced by the state-of-the-art methods. It is the first technique that delivers both the best explanations and highest accuracy for multiple targets simultaneously.

References

Aletras, N.; and Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, 13–22.

Antognini, D.; and Faltings, B. 2020. HotelRec: a Novel Very Large-Scale Hotel Recommendation Dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 4917–4923. Marseille, France: European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.605.

Antognini, D.; Musat, C.; and Faltings, B. 2020. Interacting with Explanations through Critiquing. URL https://arxiv.org/abs/2005.11067.

Bahdanau, D.; Cho, K.; and Bengio, Y. 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*. URL http://arxiv.org/abs/1409.0473.

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1903–1913. Brussels, Belgium. doi:10.18653/v1/D18-1216. URL https://www.aclweb.org/anthology/D18-1216.

Bouma, G. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In Chiarcos, C.; de Castilho, E.; and Stede, M., eds., *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, 31–40. Tübingen: Gunter Narr Verlag.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. 2019. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing Systems*, 10055–10065.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. S. 2020. Invariant rationalization. *arXiv preprint arXiv:2003.09772* .

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *International Conference on Machine Learning*, 883–892.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.408. URL https://www.aclweb.org/anthology/2020.acl-main.408.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015a. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1606–1615. Denver, Colorado: Association for Computational Linguistics. doi:10.3115/v1/N15-1184. URL https://www.aclweb.org/anthology/N15-1184.

Faruqui, M.; Tsvetkov, Y.; Yogatama, D.; Dyer, C.; and Smith, N. A. 2015b. Sparse Overcomplete Word Vector Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1491–1500.

Figurnov, M.; Mohamed, S.; and Mnih, A. 2018. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, 441–452.

Herbelot, A.; and Vecchi, E. M. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 3543–3556.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*. URL https://openreview.net/forum?id=rkE3y85ee.

Karpathy, A.; Johnson, J.; and Li, F. 2015. Visualizing and Understanding Recurrent Networks. *CoRR* abs/1506.02078. URL http://arxiv.org/abs/1506.02078.

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

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*. URL http://arxiv.org/abs/1412.6980.

Lau, J. H.; and Baldwin, T. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, 483–487.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1011. URL https://www.aclweb.org/anthology/D16-1011.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* .

Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A Structured Self-Attentive Sentence Embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal. doi:10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 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*. URL https://openreview.net/forum?id=S1jE5L5gl.

Martins, A.; and Astudillo, R. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*.

McAuley, J.; Leskovec, J.; and Jurafsky, D. 2012. Learning Attitudes and Attributes from Multi-aspect Reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, 1020–1025. Washington, DC, USA. ISBN 978-0-7695-4905-7. doi:10.1109/ICDM.2012.110. URL http://dx.doi.org/10.1109/ICDM.2012.110.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73: 1–15.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association*

for Computational Linguistics, 2227–2237. New Orleans. doi:10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.432. URL https://www.aclweb.org/anthology/2020.acl-main.432.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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*, 1135–1144.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy. URL https://www.aclweb.org/anthology/P19-1282.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1): 1929–1958.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, S.; and Manning, C. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 90–94. Jeju Island, Korea. URL https://www.aclweb.org/anthology/P12-2018.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. 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)*, 4094–4103. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1420. URL https://www.aclweb.org/anthology/D19-1420.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 795–804. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1076. URL https://www.aclweb.org/anthology/D16-1076.

A   Appendices

A.1   Topic Words per Aspect

For each model, we computed the probability distribution of words per aspect by using the induced sub-masks $M_{a_1}, ..., M_{a_A}$ or attention values. Given an aspect a_i and a set of top-N words $\boldsymbol{w_{a_i}^N}$, the Normalized Pointwise Mutual Information (Bouma 2009) coherence score is:

$$\text{NPMI}(\boldsymbol{w_{a_i}^N}) = \sum_{j=2}^{N}\sum_{k=1}^{j-1} \frac{\log \frac{P(w_{a_i}^k, w_{a_i}^j)}{P(w_{a_i}^k)P(w_{a_i}^j)}}{-\log P(w_{a_i}^k, w_{a_i}^j)} \qquad (7)$$

Top words of coherent topics (i.e., aspects) should share a similar semantic interpretation, and thus interpretability of a topic can be estimated by measuring how many words are not related. For each aspect a_i and word w having been highlighted at least once as belonging to aspect a_i, we computed the probability $P(w|a_i)$ on each dataset and sorted them in decreasing order of $P(w|a_i)$. Unsurprisingly, we found that the most common words are stop words such as "a" and "it", because masks are mostly word sequences instead of individual words. To gain a better interpretation of the aspect words, we followed the procedure in (McAuley, Leskovec, and Jurafsky 2012): we first computed the averages across all aspect words for each word w as follows:

$$b_w = \frac{1}{|A|}\sum_{i=1}^{|A|} P(w|a_i) \qquad (8)$$

It represents a general distribution that includes words common to all aspects. The final word distribution per aspect is computed by removing the general distribution as follows:

$$\hat{P}(w|a_i) = P(w|a_i) - b_w \qquad (9)$$

After generating the final word distribution per aspect, we picked the top ten words and asked two human annotators to identify intruder words (i.e., words not matching the corresponding aspect). We show in Table 5 and Table 6 (and also Table 9 in Appendix A.2) the top ten words for each aspect, where **red** denotes all words identified as unrelated to the aspect by the two annotators. Generally, our model finds better sets of words across the three datasets compared with other methods. Additionally, we observe that the aspects can be easily recovered, given its top words.

A.2   Results Decorrelated Beer Dataset

We provide additional details of Section 5.2. Table 7 presents descriptive statistics of *Beer* and *Hotel* datasets with the *decorrelated* subset of beer reviews from (Lei, Barzilay, and Jaakkola 2016; Li, Monroe, and Jurafsky 2016; Chang et al. 2019, 2020). The results of the multi-aspect sentiment classification experiment are shown in Table 8. Samples are available in Figure 13 and Figure 14. Table 9 contains the results of the intruder task.

A.3   Visualization of the Multi-Dimensional Facets of Reviews

We randomly sampled reviews from each dataset and computed the masks and attentions of four models: our Multi-Target Masker (*MTM*), the Single-Aspect Masker (*SAM*) (Lei,

Barzilay, and Jaakkola 2016), and two attention models with additive and sparse attention, called Multi-Aspect Attentions (*MAA*) and Multi-Aspect Sparse-Attentions (*MASA*), respectively (see Section 4.3). Each color represents an aspect and the shade its confidence. All models generate soft attentions or masks besides SAM, which does hard masking. Samples for the *Beer* and *Hotel* datasets are shown in Figure 9, 10, 11, and 12, respectively.

A.4   Baseline Architectures
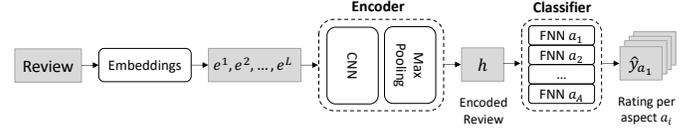
Figure 4: Baseline model Emb + Enc$_{\text{CNN}}$ + Clf (*BASE*).
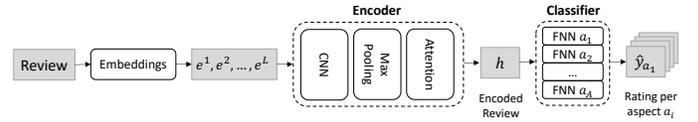
Figure 5: Baseline model Emb + Enc$_{\text{CNN}}$ + A$_{\text{Shared}}$ + Clf (*SAA*, CNN variant).
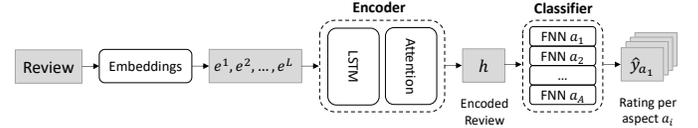
Figure 6: Baseline model Emb + Enc$_{\text{LSTM}}$ + A$_{\text{Shared}}$ + Clf (*SAA*, LSTM variant).
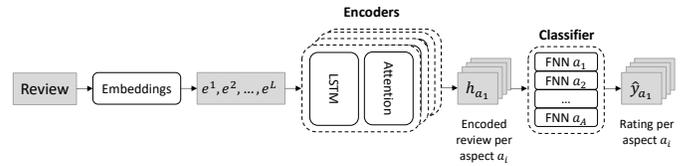
Figure 7: Baselines Emb + Enc$_{\text{LSTM}}$ + A$_{\text{Aspect-wise}}^{[\text{Sparse}]}$ + Clf. Attention is either additive (*MAA*) or sparse (*MASA*).

B   Additional Training Details

Most of the time, the model converges under 20 epochs (maximum of 20 and 3 minutes per epoch for the *Beer* and *Hotel* dataset, respectively). The range of hyperparameters are the following for MTM (similar for other models).

- Learning rate: $[0.001, 0.0005, 0.00075]$;
- Hidden size: $[50, 100, 200]$;
- Filter numbers (CNN): $[50, 100, 200]$;

	Model	Top-10 Words
Appearance	SAM	**nothing** beautiful lager nice **average** macro lagers corn **rich** gorgeous
	MASA	lacing head lace **smell** amber retention beer nice carbonation glass
	MAA	head lacing **smell aroma** color pours amber glass white retention
	MTM (Ours)	head lacing **smell** white lace retention glass **aroma** tan thin
Smell	SAM	faint **nice mild** light slight complex good wonderful grainy great
	MASA	aroma hops nose chocolate caramel malt citrus fruit smell fruits
	MAA	**taste** hints hint **lots t- starts** blend mix **upfront** malts
	MTM (Ours)	**taste** malt aroma hops sweet citrus caramel nose malts chocolate
Palate	SAM	thin **bad** light watery creamy silky medium body smooth **perfect**
	MASA	smooth light medium thin creamy **bad** watery **full** crisp **clean**
	MAA	good beer carbonation smooth **drinkable** medium bodied **nice** body **overall**
	MTM (Ours)	carbonation medium mouthfeel body smooth bodied **drinkability** creamy light **overall**
Taste	SAM	**decent** great complex delicious tasty favorite **pretty** sweet **well best**
	MASA	good **drinkable nice** tasty great enjoyable **decent solid** balanced **average**
	MAA	malt hops flavor hop flavors caramel malts bitterness bit chocolate
	MTM (Ours)	malt sweet hops flavor bitterness finish chocolate bitter caramel sweetness

Table 5: Top ten words for each aspect from the **Beer** dataset, learned by various models. **Red** denotes intruders according to two annotators. Found words are generally noisier due to the high correlation between *Taste* and other aspects. However, *MTM* provides better results than other methods.

	Model	Top-10 Words
Service	SAM	staff service friendly nice told helpful good great lovely manager
	MASA	friendly helpful told rude nice good pleasant asked enjoyed worst
	MAA	staff service helpful friendly nice good rude excellent great desk
	MTM (Ours)	staff friendly service desk helpful manager reception told rude asked
Cleanliness	SAM	clean cleaned dirty toilet smell cleaning sheets comfortable nice hair
	MASA	clean dirty cleaning spotless stains cleaned cleanliness mold filthy bugs
	MAA	clean dirty cleaned filthy stained well spotless carpet sheets stains
	MTM (Ours)	clean dirty bathroom room bed cleaned sheets smell carpet toilet
Value	SAM	good stay great well **dirty** recommend worth definitely **friendly** charged
	MASA	great good poor excellent terrible awful **dirty** horrible **disgusting comfortable**
	MAA	**night stayed** stay **nights 2 day** price **water 4 3**
	MTM (Ours)	good price expensive paid cheap worth better pay overall disappointed
Location	SAM	location close far place walking **definitely** located **stay** short view
	MASA	location beach walk hotel town located restaurants walking close taxi
	MAA	location hotel place located close far area beach view situated
	MTM (Ours)	location great area walk beach hotel town close city street
Room	SAM	**dirty clean** small best comfortable large worst modern **smell** spacious
	MASA	comfortable small spacious nice large dated well tiny modern basic
	MAA	room rooms bathroom bed spacious small beds large shower modern
	MTM (Ours)	comfortable room small spacious nice modern rooms large tiny walls

Table 6: Top ten words for each aspect from the **Hotel** dataset, learned by various models. **Red** denotes intruders according to human annotators. Besides *SAM*, all methods find similar words for most aspects except the aspect *Value*. The top words of *MTM* do not contain any intruder.

- Bi-directional (LSTM): $[True, False]$;
- Dropout: $[0, 0.1, 0.2]$;
- Weight decay: $[0, 1e^{-6}, 1e^{-8}, 1e^{-10}]$;
- Gumbel Temperature τ: $[0.5, 0.8, 1.0, 1.2]$;
- λ_{sel}: $[0.01, 0.02, 0.03, 0.04, 0.05]$;

- λ_p: $[0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15]$;
- λ_{cont}: $[0.02, 0.04, 0.06, 0.08, 0.10]$;

We used a computer with the following configuration: 2x Intel Xeon E5-2680, 256GB RAM, 1x Nvidia Titan X, Ubuntu 18.04, Python 3.6, PyTorch 1.3.1, CUDA 9.2.

Dataset	Beer	Hotel	Decorrelated Beer
Number of reviews	$1,586,259$	$140,000$	$280,000$
Average word-length of review	147.1 ± 79.7	188.3 ± 50.0	157.5 ± 84.3
Average sentence-length of review	10.3 ± 5.4	10.4 ± 4.4	11.0 ± 5.7
Number of aspects	4	5	3
Average ratio of $\oplus$ over $\ominus$ reviews per aspect	12.89	1.02	3.29
Average correlation between aspects	**71.8**%	**63.0**%	**27.2**%
Max correlation between two aspects	**73.4**%	**86.5**%	**29.8**%

Table 7: Statistics of the multi-aspect review datasets. *Beer* and *Hotel* represent real-world beer and hotel reviews, respectively. *Decorrelated Beer* is a subset of the *Beer* dataset with a low-correlation assumption between aspect ratings, leading to a more straightforward and unrealistic dataset.

Interp.		Model	Params	F1 Score			
				Macro	A_1	A_2	A_3
None	SENT	Sentiment Majority	$426k$	68.89	67.48	73.49	65.69
	BASE	$\text{Emb}_{200} + \text{Enc}_{\text{CNN}} + \text{Clf}$	$173k$	78.23	78.38	80.86	75.47
Coarse-grained	SAA	$\text{Emb}_{200} + \text{Enc}_{\text{CNN}} + \text{A}_{\text{Shared}} + \text{Clf}$	$196k$	78.19	77.43	80.96	76.16
		$\text{Emb}_{200} + \text{Enc}_{\text{LSTM}} + \text{A}_{\text{Shared}} + \text{Clf}$	$186k$	78.16	75.88	81.25	77.36
Fine-grained	NB-SVM	(Wang and Manning 2012)	$3 \cdot 426k$	74.60	73.50	77.32	72.99
	SAM	(Lei, Barzilay, and Jaakkola 2016)	$3 \cdot 644k$	77.06	77.36	78.99	74.83
	MASA	$\text{Emb}_{200} + \text{Enc}_{\text{LSTM}} + \text{A}_{\text{Aspect-wise}}^{\text{Sparse}} + \text{Clf}$	$458k$	78.82	77.35	81.65	77.47
	MAA	$\text{Emb}_{200} + \text{Enc}_{\text{LSTM}} + \text{A}_{\text{Aspect-wise}} + \text{Clf}$	$458k$	78.96	78.54	81.56	76.79
	MTM	$\text{Emb}_{200} + \text{Masker} + \text{Enc}_{\text{CNN}} + \text{Clf}$ (Ours)	$274k$	79.32	78.58	81.71	77.66
	MTM$^{\text{C}}$	**$\text{Emb}_{200+4} + \text{Enc}_{\text{CNN}} + \text{Clf}$ (Ours)**	$175k$	**79.66**	**78.74**	**82.02**	**78.22**

Table 8: Performance of the multi-aspect sentiment classification task for the ***decorrelated** Beer* dataset.

	Model	Top-10 Words
Appearance	SAM	head color white brown dark lacing **pours** amber clear black
	MASA	head lacing lace retention glass foam color amber yellow cloudy
	MAA	nice dark amber **pours** black hazy brown **great** cloudy clear
	MTM (Ours)	head color lacing white brown clear amber glass black retention
Smell	SAM	sweet malt hops coffee chocolate citrus hop strong smell aroma
	MASA	smell aroma nose smells sweet aromas scent hops malty roasted
	MAA	**taste** smell aroma sweet chocolate **lacing** malt roasted hops nose
	MTM (Ours)	smell aroma nose smells sweet malt citrus chocolate caramel aromas
Palate	SAM	mouthfeel smooth medium carbonation bodied watery body thin creamy **full**
	MASA	mouthfeel medium smooth body **nice** m- feel bodied mouth **beer**
	MAA	carbonation mouthfeel medium **overall** smooth finish body **drinkability** bodied watery
	MTM (Ours)	mouthfeel carbonation medium smooth body bodied **drinkability** good mouth thin

Table 9: Top ten words for each aspect from the ***decorrelated** Beer* dataset, learned by various models. **Red** denotes intruders according to two annotators. For the three aspects, MTM has only one word considered as an intruder, followed by MASA with SAM (two) and MAA (six).

C   Applying MTM for Image Classification

Our method can be applied to different types of inputs like images. In that case, the words in the text data are replaced by regions or patches in an image, obtained by a classical pre-trained convolutional neural network. The sparsity constraint stays similar, and the continuity is based on 2D distances

instead. We will investigate this type of data in future work.

D   Impact of the Length of the Rationales

We are interested in studying 1. the distribution of the rationales' lengths, and 2. the effect of the length in terms of the precision, recall, and F1 score. In Section 5.1, we report only

the precision on three aspects at a similar number of selected words, for a fair comparison with prior work, on the human sentence-level aspect annotations.

We use the *Beer* dataset and also consider the extra aspect *Overall* (which leads to five aspects in total), and we train MTM on it as in Section 4.1. We compute the distribution of the explanation length (in percent) on the validation set for each aspect $a_i, i = 1, ..., 5$. We calculate all the 100 percentiles $P_p^{a_i}$. Then, we infer the sub-masks M_{a_i} and generate the rationale by selecting each word x^ℓ, whose relevance towards a_i satisfies $P(m_{a_i}^\ell | x^\ell) \geq P_p^{a_i}$. Finally, we compute the precision, recall, and F1 score on the human annotation for each percentile $P_p^{a_i}$.

Figure 8 shows the result of the five aspects. First, we observe that the length distribution of the rationales for the aspect *Overall* is the most spread compared to the other aspects, and these of the aspect *Smell* the least; the three left aspects share a similar distribution.

In terms of precision, we notice that the aspect *Palate* drops quickly compared to the other ones that decrease linearly when we augment the portion of selected text. According to RateBeer[7], the aspect *Palate* is the most difficult one to rate, which confirms our findings. For most aspects the precision remains high after selecting 75% of the text ($5 \cdot 15\%$), showing the effectiveness of our approach. In terms of recall, they all increase unsurprisingly when highlighted more words.

Finally, we show that MTM achieves a very high precision ($> 90\%$) by highlighting only 25% of the document, that reduces the cognitive load of a user to identify the important parts of the document.
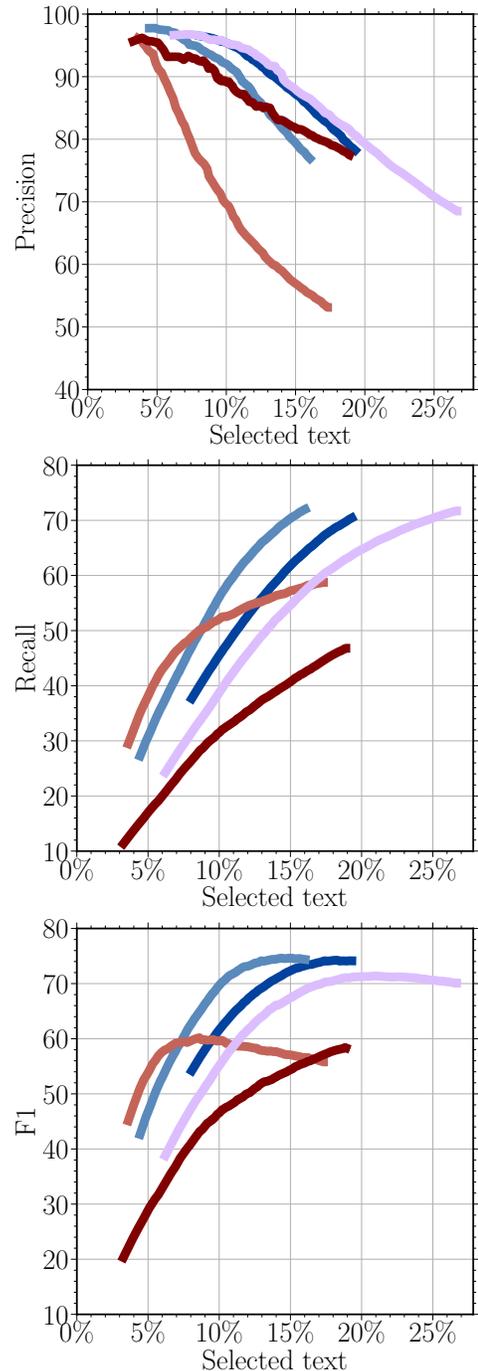

Figure 8: Performance related to human evaluation, showing the precision, recall, and F1 scores of the rationales for each aspect (with the overall rating) of the *Beer* dataset. The percentage of words indicates the number of highlighted words of the full review.

Appearance Smell Palate Taste

Multi-Target Masker (Ours)

75cl bottle shared with larrylsb , pre - grad . bright , hazy gold with a big white head . the flavor has bursting fruit and funky yeast with tropical and peach standing out . the flavor has the same intense fruitiness , with a funky , lightly tart edge , and a nice hop balance . dry and refreshing on the tongue . medium bodied with perfect carbonation that livens up the palate . this was just beautiful stuff that i 'm already craving more of .

Appearance Smell Palate Taste

Single-Aspect Masker (Lei et al., 2016)

75cl bottle shared with larrylsb , pre - grad . bright , hazy gold with a big white head . the flavor has bursting fruit and funky yeast with tropical and peach standing out . the flavor has the same intense fruitiness , with a funky , lightly tart edge , and a nice hop balance . dry and refreshing on the tongue . medium bodied with perfect carbonation that livens up the palate . this was just beautiful stuff that i 'm already craving more of .

Appearance Smell Palate Taste

Multi-Aspect Attentions

75cl bottle shared with larrylsb , pre - grad . bright , hazy gold with a big white head . the flavor has bursting fruit and funky yeast with tropical and peach standing out . the flavor has the same intense fruitiness , with a funky , lightly tart edge , and a nice hop balance . dry and refreshing on the tongue . medium bodied with perfect carbonation that livens up the palate . this was just beautiful stuff that i 'm already craving more of .

Appearance Smell Palate Taste

Multi-Aspect Sparse-Attentions

75cl bottle shared with larrylsb , pre - grad . bright , hazy gold with a big white head . the flavor has bursting fruit and funky yeast with tropical and peach standing out . the flavor has the same intense fruitiness , with a funky , lightly tart edge , and a nice hop balance . dry and refreshing on the tongue . medium bodied with perfect carbonation that livens up the palate . this was just beautiful stuff that i 'm already craving more of .

Figure 9: A sample review from the *Beer* dataset, with computed masks from different methods. *MTM* achieves near-perfect annotations, while *SAM* highlights only two words where one is ambiguous with respect to the four aspects. *MAA* mixes between the aspect *Appearance* and *Smell*. *MASA* identifies some words but lacks coverage.

Appearance Smell Palate Taste

Multi-Target Masker (Ours)

sa 's harvest pumpkin ale 2011 . had this last year , loved it , and bought 6 harvest packs and saved the pumpkins and the dunkel 's ... not too sure why sa dropped the dunkel , i think it would make a great standard to them . pours a dark brown with a 1 " bone white head , that settles down to a thin lace across the top of the brew . smells of the typical pumpkin pie spice , along with a good squash note . tastes just like last years , very subtle , nothing over the top . a damn good pumpkin ale that is worth seeking out . when i mean everything is subtle i mean everything . nothing is overdone in this pumpkin ale , and is a great representation of the original style . mouthfeel is somewhat thick , with a pleasant coating feel . overall , i loved it last year , and i love it this year . do n't get me wrong , its no pumpking , but this is a damn fine pumpkin ale that could hold its own any day among all the others . i would rate this as my 4th favorite pumpkin ale to date . i 'm not sure why the bros rated it so low , but do n't take their opinion , make your own !

Appearance Smell Palate Taste

Single-Aspect Masker (Lei et al., 2016)

sa 's harvest pumpkin ale 2011 . had this last year , loved it , and bought 6 harvest packs and saved the pumpkins and the dunkel 's ... not too sure why sa dropped the dunkel , i think it would make a great standard to them . pours a dark brown with a 1 " bone white head , that settles down to a thin lace across the top of the brew . smells of the typical pumpkin pie spice , along with a good squash note . tastes just like last years , very subtle , nothing over the top . a damn good pumpkin ale that is worth seeking out . when i mean everything is subtle i mean everything . nothing is overdone in this pumpkin ale , and is a great representation of the original style . mouthfeel is somewhat thick , with a pleasant coating feel . overall , i loved it last year , and i love it this year . do n't get me wrong , its no pumpking , but this is a damn fine pumpkin ale that could hold its own any day among all the others . i would rate this as my 4th favorite pumpkin ale to date . i 'm not sure why the bros rated it so low , but do n't take their opinion , make your own !

Appearance Smell Palate Taste

Multi-Aspect Attentions

sa 's harvest pumpkin ale 2011 . had this last year , loved it , and bought 6 harvest packs and saved the pumpkins and the dunkel 's ... not too sure why sa dropped the dunkel , i think it would make a great standard to them . pours a dark brown with a 1 " bone white head , that settles down to a thin lace across the top of the brew . smells of the typical pumpkin pie spice , along with a good squash note . tastes just like last years , very subtle , nothing over the top . a damn good pumpkin ale that is worth seeking out . when i mean everything is subtle i mean everything . nothing is overdone in this pumpkin ale , and is a great representation of the original style . mouthfeel is somewhat thick , with a pleasant coating feel . overall , i loved it last year , and i love it this year . do n't get me wrong , its no pumpking , but this is a damn fine pumpkin ale that could hold its own any day among all the others . i would rate this as my 4th favorite pumpkin ale to date . i 'm not sure why the bros rated it so low , but do n't take their opinion , make your own !

Appearance Smell Palate Taste

Multi-Aspect Sparse-Attentions

sa 's harvest pumpkin ale 2011 . had this last year , loved it , and bought 6 harvest packs and saved the pumpkins and the dunkel 's ... not too sure why sa dropped the dunkel , i think it would make a great standard to them . pours a dark brown with a 1 " bone white head , that settles down to a thin lace across the top of the brew . smells of the typical pumpkin pie spice , along with a good squash note . tastes just like last years , very subtle , nothing over the top . a damn good pumpkin ale that is worth seeking out . when i mean everything is subtle i mean everything . nothing is overdone in this pumpkin ale , and is a great representation of the original style . mouthfeel is somewhat thick , with a pleasant coating feel . overall , i loved it last year , and i love it this year . do n't get me wrong , its no pumpking , but this is a damn fine pumpkin ale that could hold its own any day among all the others . i would rate this as my 4th favorite pumpkin ale to date . i 'm not sure why the bros rated it so low , but do n't take their opinion , make your own !

Figure 10: A sample review from the *Beer* dataset, with computed masks from different methods. *MTM* can accurately identify what parts of the review describe each aspect. *MAA* provides very noisy labels due to the high imbalance and correlation between aspects, while *MASA* highlights only a few important words. We can see that *SAM* is confused and performs a poor selection.

Service Cleanliness Value Location Room

Multi-Target Masker (Ours)

i stayed at daulsol in september 2013 and could n't have asked for anymore for the price ! ! it is a great location .... only 2 minutes walk to jet , space and sankeys with a short drive to ushuaia . the hotel is basic but cleaned daily and i did nt have any problems at all with the bathroom or kitchen facilities . the lady at reception was really helpful and explained everything we needed to know ..... even when we managed to miss our flight she let us stay around and use the facilities until we got on a later flight . there are loads of restaurants in the vicinity and supermarkets and shops right outside . i loved these apartments so much that i booked to come back for september 2014 ! ! can not wait :)

Service Cleanliness Value Location Room

Single-Aspect Masker (Lei et al., 2016)

i stayed at daulsol in september 2013 and could n't have asked for anymore for the price ! ! it is a great location .... only 2 minutes walk to jet , space and sankeys with a short drive to ushuaia . the hotel is basic but cleaned daily and i did nt have any problems at all with the bathroom or kitchen facilities . the lady at reception was really helpful and explained everything we needed to know ..... even when we managed to miss our flight she let us stay around and use the facilities until we got on a later flight . there are loads of restaurants in the vicinity and supermarkets and shops right outside . i loved these apartments so much that i booked to come back for september 2014 ! ! can not wait :)

Service Cleanliness Value Location Room

Multi-Aspect Attentions

i stayed at daulsol in september 2013 and could n't have asked for anymore for the price ! ! it is a great location .... only 2 minutes walk to jet , space and sankeys with a short drive to ushuaia . the hotel is basic but cleaned daily and i did nt have any problems at all with the bathroom or kitchen facilities . the lady at reception was really helpful and explained everything we needed to know ..... even when we managed to miss our flight she let us stay around and use the facilities until we got on a later flight . there are loads of restaurants in the vicinity and supermarkets and shops right outside . i loved these apartments so much that i booked to come back for september 2014 ! ! can not wait :)

Service Cleanliness Value Location Room

Multi-Aspect Sparse-Attentions

i stayed at daulsol in september 2013 and could n't have asked for anymore for the price ! ! it is a great location .... only 2 minutes walk to jet , space and sankeys with a short drive to ushuaia . the hotel is basic but cleaned daily and i did nt have any problems at all with the bathroom or kitchen facilities . the lady at reception was really helpful and explained everything we needed to know ..... even when we managed to miss our flight she let us stay around and use the facilities until we got on a later flight . there are loads of restaurants in the vicinity and supermarkets and shops right outside . i loved these apartments so much that i booked to come back for september 2014 ! ! can not wait :)

Figure 11: A sample review from the *Hotel* dataset, with computed masks from different methods. *MTM* emphasizes consecutive words, identifies essential spans while having a small amount of noise. *SAM* focuses on certain specific words and spans, but labels are ambiguous. The *MAA* model highlights many words, ignores a few crucial key-phrases, but labels are noisy when the confidence is low. *MASA* provides noisier tags than *MAA*.

Multi-Target Masker (Ours)

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! i went to the end of the street one night and got mugged .. all my money , camera .. everything ! got sratches on my chest which has now scarred me , and i had bruises at the time . just make sure you have got someone with you at all times , the local people are very renound for this , went to police station the next day (in old town) and there was many english in there reporting their muggings from the day before . shocking ! ! apart from this incident (on the first night we arrived :( ) we had a good time in the end , plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Service Cleanliness Value Location Room

Single-Aspect Masker (Lei et al., 2016)

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! i went to the end of the street one night and got mugged .. all my money , camera .. everything ! got sratches on my chest which has now scarred me , and i had bruises at the time . just make sure you have got someone with you at all times , the local people are very renound for this . went to police station the next day (in old town) and there was many english in there reporting their muggings from the day before . shocking ! ! apart from this incident (on the first night we arrived :( ) we had a good time in the end , plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Service Cleanliness Value Location Room

Multi-Aspect Attentions

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! i went to the end of the street one night and got mugged .. all my money , camera .. everything ! got sratches on my chest which has now scarred me , and i had bruises at the time . just make sure you have got someone with you at all times , the local people are very renound for this . went to police station the next day (in old town) and there was many english in there reporting their muggings from the day before . shocking ! ! apart from this incident (on the first night we arrived :( ) we had a good time in the end , plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Service Cleanliness Value Location Room

Multi-Aspect Sparse-Attentions

stayed at the parasio 10 apartments early april 2011 . reception staff absolutely fantastic , great customer service .. ca nt fault at all ! we were on the 4th floor , facing the front of the hotel .. basic , but nice and clean . good location , not too far away from the strip and beach (10 min walk) . however .. do not go out alone at night at all ! i went to the end of the street one night and got mugged .. all my money , camera .. everything ! got sratches on my chest which has now scarred me , and i had bruises at the time . just make sure you have got someone with you at all times , the local people are very renound for this . went to police station the next day (in old town) and there was many english in there reporting their muggings from the day before . shocking ! ! apart from this incident (on the first night we arrived :( ) we had a good time in the end , plenty of laughs and everything is very cheap ! beer - 1euro ! fryups - 2euro . would go back again , but maybe stay somewhere else closer to the beach (sol pelicanos etc) .. this hotel is next to an alley called ' muggers alley '

Figure 12: A sample review from the *Hotel* dataset, with computed masks from different methods. Our *MTM* model finds most of the crucial span of words with a small amount of noise. *SAM* lacks coverage but identifies words where half are correctly tags and the others ambiguous. *MAA* partially correctly highlights words for the aspects *Service*, *Location*, and *Value* while missing out on the aspect *Cleanliness*. *MASA* confidently finds a few important words.
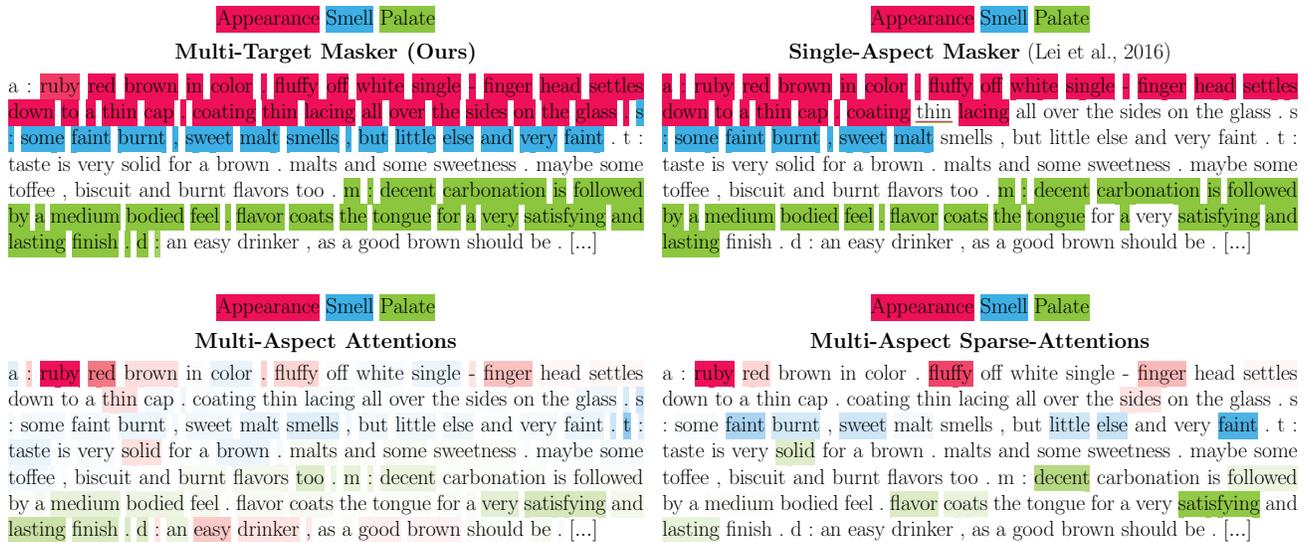
Multi-Target Masker (Ours)

a : ruby red brown in color . fluffy off white single - finger head settles down to a thin cap . coating thin lacing all over the sides on the glass . s : some faint burnt , sweet malt smells , but little else and very faint . t : taste is very solid for a brown . malts and some sweetness . maybe some toffee , biscuit and burnt flavors too . m : decent carbonation is followed by a medium bodied feel . flavor coats the tongue for a very satisfying and lasting finish . d : an easy drinker , as a good brown should be . [...]

Single-Aspect Masker (Lei et al., 2016)

a : ruby red brown in color . fluffy off white single - finger head settles down to a thin cap . coating thin lacing all over the sides on the glass . s : some faint burnt , sweet malt smells , but little else and very faint . t : taste is very solid for a brown . malts and some sweetness . maybe some toffee , biscuit and burnt flavors too . m : decent carbonation is followed by a medium bodied feel . flavor coats the tongue for a very satisfying and lasting finish . d : an easy drinker , as a good brown should be . [...]

Multi-Aspect Attentions

a : ruby red brown in color . fluffy off white single - finger head settles down to a thin cap . coating thin lacing all over the sides on the glass . s : some faint burnt , sweet malt smells , but little else and very faint . t : taste is very solid for a brown . malts and some sweetness . maybe some toffee , biscuit and burnt flavors too . m : decent carbonation is followed by a medium bodied feel . flavor coats the tongue for a very satisfying and lasting finish . d : an easy drinker , as a good brown should be . [...]

Multi-Aspect Sparse-Attentions

a : ruby red brown in color . fluffy off white single - finger head settles down to a thin cap . coating thin lacing all over the sides on the glass . s : some faint burnt , sweet malt smells , but little else and very faint . t : taste is very solid for a brown . malts and some sweetness . maybe some toffee , biscuit and burnt flavors too . m : decent carbonation is followed by a medium bodied feel . flavor coats the tongue for a very satisfying and lasting finish . d : an easy drinker , as a good brown should be . [...]

Figure 13: A sample review from the ***decorrelated*** *Beer* dataset, with computed masks from different methods. Our model *MTM* highlights all the words corresponding to the aspects. *SAM* only highlights the most crucial information, but some words are missing out and one is ambiguous. *MAA* and *MASA* fail to identify most of the words related to the aspect *Appearance*, and only a few words have high confidence, resulting in noisy labeling. Additionally, *MAA* considers words belonging to the aspect *Taste* whereas this dataset does not include it in the aspect set (because it has a high correlation with other rating scores).
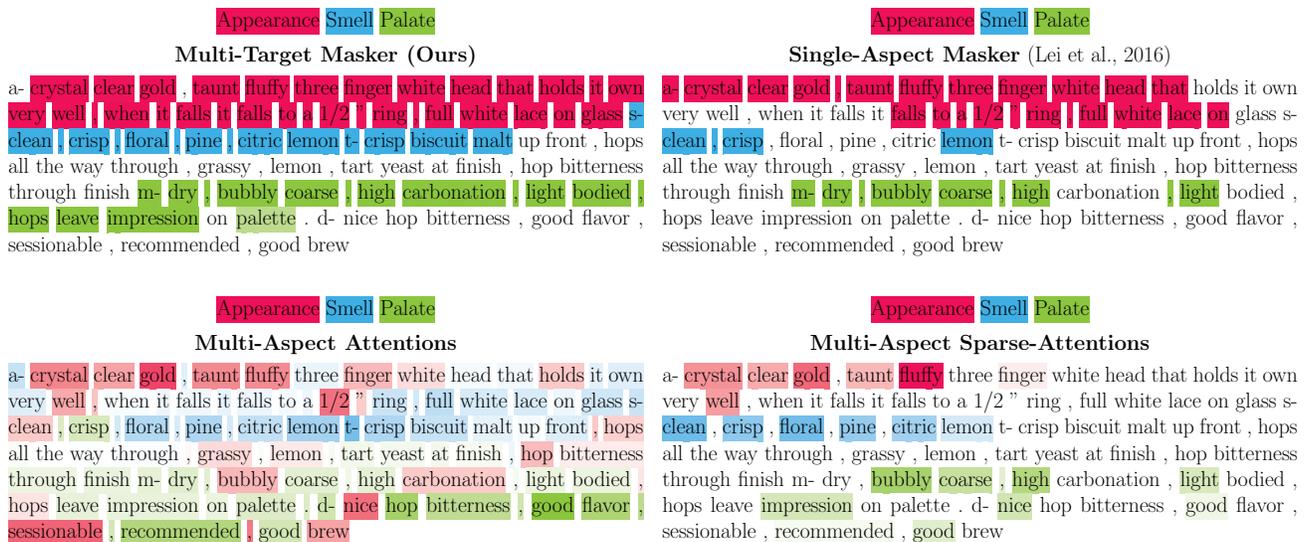
Multi-Target Masker (Ours)

a- crystal clear gold , taunt fluffy three finger white head that holds it own very well , when it falls it falls to a 1/2 " ring , full white lace on glass s- clean , crisp , floral , pine , citric lemon t- crisp biscuit malt up front , hops all the way through , grassy , lemon , tart yeast at finish , hop bitterness through finish m- dry , bubbly coarse , high carbonation , light bodied , hops leave impression on palette . d- nice hop bitterness , good flavor , sessionable , recommended , good brew

Single-Aspect Masker (Lei et al., 2016)

a- crystal clear gold , taunt fluffy three finger white head that holds it own very well , when it falls it falls to a 1/2 " ring , full white lace on glass s- clean , crisp , floral , pine , citric lemon t- crisp biscuit malt up front , hops all the way through , grassy , lemon , tart yeast at finish , hop bitterness through finish m- dry , bubbly coarse , high carbonation , light bodied , hops leave impression on palette . d- nice hop bitterness , good flavor , sessionable , recommended , good brew

Multi-Aspect Attentions

a- crystal clear gold , taunt fluffy three finger white head that holds it own very well , when it falls it falls to a 1/2 " ring , full white lace on glass s- clean , crisp , floral , pine , citric lemon t- crisp biscuit malt up front , hops all the way through , grassy , lemon , tart yeast at finish , hop bitterness through finish m- dry , bubbly coarse , high carbonation , light bodied , hops leave impression on palette . d- nice hop bitterness , good flavor , sessionable , recommended , good brew

Multi-Aspect Sparse-Attentions

a- crystal clear gold , taunt fluffy three finger white head that holds it own very well , when it falls it falls to a 1/2 " ring , full white lace on glass s- clean , crisp , floral , pine , citric lemon t- crisp biscuit malt up front , hops all the way through , grassy , lemon , tart yeast at finish , hop bitterness through finish m- dry , bubbly coarse , high carbonation , light bodied , hops leave impression on palette . d- nice hop bitterness , good flavor , sessionable , recommended , good brew

Figure 14: A sample review from the ***decorrelated*** *Beer* dataset, with computed masks from different methods. *MTM* finds the exact parts corresponding to the aspect *Appearance* and *Palate* while covering most of the aspect *Smell*. *SAM* identifies key-information without any ambiguity, but lacks coverage. *MAA* highlights confidently nearly all the words while having some noise for the aspect *Appearance*. *MASA* selects confidently only most predictive words.