Multi-Dimensional Explanation of Ratings from Reviews

Diego Antognini
Artificial Intelligence Laboratory
EPFL, Switzerland
diego.antognini@epfl.ch

Claudiu Musat
Swisscom
claudiu.musat@swisscom.com

Boi Faltings
Artificial Intelligence Laboratory
EPFL, Switzerland
boi.faltings@epfl.ch

Abstract

Automated predictions require explanations to be interpretable by humans. In the review domain, finding the justifications of aspect ratings brings a fine-grained understanding of opinions on product features and users' preferences. Past work used attention and other mechanisms to find words that predict the sentiment to an aspect, but often they result in a tradeoff between noisy explanations or a drop in accuracy. In this paper, we propose Multi-Aspect Masker (MAM), a multi-task learning neural model that generates, in an unsupervised manner, a *multi-dimensional mask* that justifies aspect ratings of reviews. The novelty lies in the regularization that guides MAM to induce long and meaningful explanations per aspect. Evaluation using standard metrics and human annotations show that the resulting masks are more accurate and coherent than those generated by state-of-the-art methods. Experiments on two datasets show that MAM is the first to bring the best of both worlds. It infers interpretable masks that embed high-quality representations and improves F1 scores of multi-aspect sentiment classification.

1 Introduction

Neural models have become the standard for many natural language processing tasks. Despite the significant performance gains achieved by these complex models, they offer little *transparency* concerning their inner workings. Thus, they come at the cost of *interpretability* (Jain and Wallace, 2019).

In many domains, automated predictions have a real *impact* on the final decision, such as treatments in the field of medicine. Thus, it is important to provide the underlying reasons for such a decision. We claim that integrating interpretability into a (neural) model should supply reasoning for the prediction and yield better performance. However, *explaining* a prediction is an ambiguous and challenging goal.

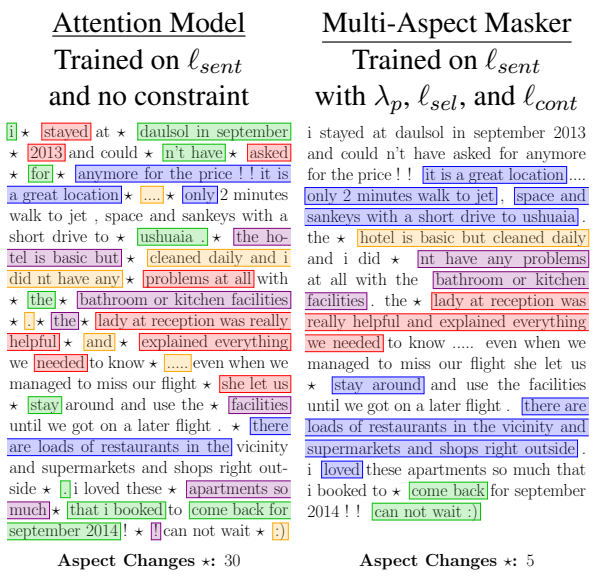

Figure 1: A hotel review with explanations produced by an attention model and our Multi-Aspect Masker model. The colors represent the aspects: Room, Value, Service, Cleanliness, and Location. Masks lead to mostly long sequences describing clearly each aspect (one switch ⋆ per aspect), while attention to many short, noisy, and interleaving sequences (30 changes ⋆ among aspects). Highlighted words correspond to the highest aspect-attention scores above $1/L$ (i.e., a uniform distribution), and the aspect a_i maximizing $P(m^{\ell}_{a_i}|x^{\ell})$.

Prior work includes various methods that find the explanation in the input text — called rationale or mask of a target variable. A *mask*[1] is defined as one or multiple snippets from the input text that are alone sufficient for the prediction (Lei et al., 2016).

Many works have been applied to single-aspect sentiment analysis for reviews, where the *ambiguity* of what is meant by an explanation is minimal. In this case, we define an aspect as an attribute of a product or service, such as *Location* or *Cleanliness*

[1] In the rest of the paper, we will use the terms mask, explanation, rationale, and justification interchangeably.

for the hotel domain. Three different methods exist to infer masks: using reinforcement learning and a trained model (Li et al., 2016), generating masks in an unsupervised manner and jointly with the loss function (Lei et al., 2016), or including annotations during training (Bao et al., 2018).

On the one hand, a strict assignment of words to aspects might lead to ambiguities that are difficult to capture with a *binary mask*. In the text *"The room was large, clean and close to the beach."*, *"room"* refers to the aspects *Room*, *Cleanliness* and *Location*. On the other, collecting human-provided rationales at scale is expensive and impractical.

In this work, we study interpretable multi-aspect sentiment classification. We describe an architecture for generating a *probabilistic (soft) multi-dimensional mask* (one dimension per aspect), in an unsupervised and multi-task learning manner, and predicting the sentiment of *multiple* aspects jointly. We show that the induced mask (Figure 1 right) is (1) beneficial for identifying which parts of the review relate to which aspects, and (2) for capturing ambiguities of words belonging to multiple aspects. Thus, the induced mask provides fine-grained interpretability and improves the final performance.

Traditionally interpretability came at the cost of reduced performance. In contrast, our evaluation shows that, on two datasets in the beer and hotel domain, our model outperforms strong baselines 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, and (2) further improve the sentiment prediction for all aspects. Therefore, interpretability does not come at a cost in our paradigm. The contributions of this work can be summarized as follows:

- We propose a Multi-Aspect Masker (*MAM*): an end-to-end neural model for multi-aspect sentiment classification that provides fine-grained interpretability jointly. Given a text review as input, the model predicts the sentiments of multiple aspects and highlights long sequences of words explaining the current predicted rating for each aspect;
- We show that interpretability does not come at a cost: MAM significantly outperforms attention and strong baselines, both in terms of mask precision, coherence, and aspect rating performance. Furthermore, the level of interpretability is controllable by two regularizers;
- Finally, we release a new dataset for multi-aspect sentiment classification: 140k reviews from TripAdvisor, each with five aspects and their corresponding ratings.[2]

2 Related Work

2.1 Interpretability

Developing interpretable models is of considerable interest to the broader research community, even more pronounced with neural models (Doshi-Velez and Kim, 2017). There has been much work with a multitude of approaches: Montavon et al. (2018) analyzed and visualized state activation, Herbelot and Vecchi (2015) learned sparse and interpretable word vectors. Jain and Wallace (2019) analyzed attention models. Our work differs from these approaches in terms of what is meant by an explanation. Our system identifies one or multiple short and coherent text fragments that – as a substitute of the input text – are sufficient for the predictions.

2.2 Attention-based models

Attention models (Vaswani 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 et al., 2015), rather than hard (Luong et al., 2015) or sparse ones (Martins and Astudillo, 2016). According to Jain and Wallace (2019); Serrano and Smith (2019), standard attention modules noisily predict input importance: the weights cannot provide safe and meaningful explanations. Our approach differs in two ways from attention mechanisms: the loss includes two regularizers to favor long word sequences for interpretability; the normalization is not done over the sequence length but over the aspect set for each word: each has a probability distribution over the aspects.

2.3 Multi-Aspect Sentiment Classification

Multi-aspect sentiment classification is sometimes seen as a sub-problem of sentiment analysis. McAuley et al. (2012); Pappas and Popescu-Belis (2014) employed heuristic-based methods or topic models. Neural models achieve significant improvements with less feature engineering. Yin et al. (2017) built a hierarchical attention model with aspect representations by using a set of manually defined topics. Li et al. (2018) extended this work with user attention and additional features such as

[2]We will make the code and data available.

overall rating, aspect, and user embeddings. The disadvantage of these methods is their limited interpretability because they rely on many features in addition to the review text.

2.4 Rationale-Based Models

The idea of including human rationales during training is explored in Bao et al. (2018). Although they have been shown to be beneficial, they are expensive to collect and might vary across annotators. In our work, no annotation is used.

The work most closely related to ours is Li et al. (2016) and Lei et al. (2016). Both generate *hard* rationales and address *single-aspect* sentiment classification. Their model must be trained *separately* for each aspect, which leads to ambiguities. Li et al. (2016) developed a post-training method that removes words from a review text until a trained model changes its prediction. Lei et al. (2016) provided a model that learns an aspect sentiment and its rationale jointly. However, it hinders the performance and relies on assumptions on the data, such as a small correlation between aspect ratings.

In contrast, our model (1) supports *multi-aspect* sentiment classification, (2) generates *soft multi-dimensional* masks in a *single* training; (3) the masks provide interpretability and improve the performance significantly.

3 Method: Multi-Aspect Masker (MAM)

Let X be a review composed of L words $x^1, x^2, ..., x^L$ and Y the target A-dimensional sentiment vector, corresponding to the different rated aspects. Our proposed model, called Multi-Aspect Masker (MAM), is composed of three components: 1) a *Masker* module that computes a probability distribution over the aspect set for each word, resulting in $A+1$ different masks (including one for the *not-aspect* case); 2) an *Encoder* that learns a representation of a review conditioned on the induced masks; 3) a *Classifier* that predicts the target variables. The overall model architecture is shown in Figure 2. **Our framework generalizes for other tasks, and each neural module is interchangeable with other models**.

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 convolutional network because it led to a smaller model, faster training, and performed empirically similarly to recurrent and attention models.

Let a_i denote the i^{th} aspect for $i = 1, ..., A$, and a_0 the *not-aspect* case, because many words can be irrelevant to every aspect. We define $M^\ell \in \mathbb{R}^{(A+1)}$, the aspect distribution of the input word x^ℓ as:

$$P(\mathbf{M}|X) = \prod_{\ell=1}^{L} P(M^\ell|x^\ell) = \prod_{\ell=1}^{L}\prod_{i=0}^{A} P(m^\ell_{a_i}|x^\ell)$$

Because we have categorical distributions, we cannot directly sample $P(M^\ell|x^\ell)$ and back-propagate the gradient through this discrete generation process. Instead, we model the variable $m^\ell_{a_i}$ using the Straight Through Gumbel Softmax (Jang et al., 2017; Maddison et al., 2017), to approximate sampling from a categorical distribution. 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 ℓ^{th} 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.

Compared to attention mechanisms, the word importance is a probability distribution over the targets $\sum_{t=0}^{T} P(m^\ell_{a_t}|x^\ell) = 1$ instead of a normalization over the sequence length $\sum_{\ell=1}^{L} P(a^\ell|x^\ell) = 1$.

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

$$M_{a_i} = P(m^1_{a_i}|x^1), P(m^2_{a_i}|x^2), ..., P(m^L_{a_i}|x^L)$$

To integrate the word importance from the induced sub-masks M_{a_i} within the model, we weight the word embeddings by their importance towards an aspect a_i, such that $E_{a_i} = E \odot M_{a_i} = e_1 \cdot P(m^1_{a_i}|x^1), e_2 \cdot P(m^2_{a_i}|x^2), ..., e_L \cdot P(m^L_{a_i}|x^L)$. Thereafter, each modified embedding E_{a_i} is fed into the *Encoder* block. Note that E_{a_0} is ignored because M_{a_0} only serves to absorb probabilities of words that are insignificant to every aspect.[3]

The *Encoder* includes a convolutional network, for the same reasons as stated earlier, followed by a max-over-time pooling to obtain a fixed-length feature vector. It produces the hidden representation h_{a_i} for each aspect a_i. To exploit commonalities and differences among aspects, we share the

[3] if $P(m^\ell_{a_0}|x^\ell) \approx 1.0$, it implies $\sum_{i=1}^{A} P(m^\ell_{a_i}|x^\ell) \approx 0$ and consequently, $e^\ell_{a_i} \approx \vec{0}$.

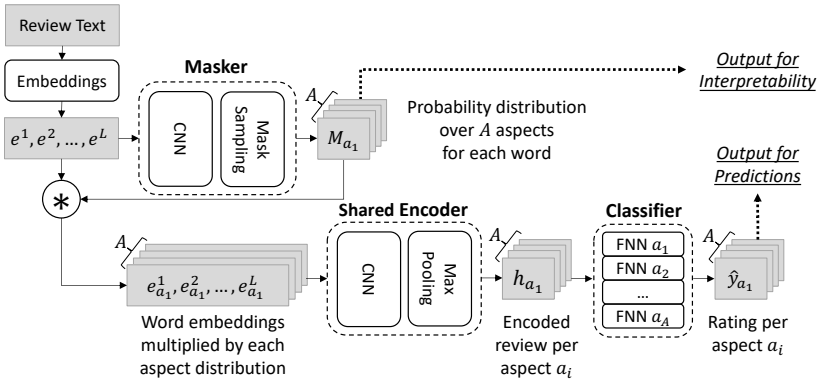

Figure 2: The proposed Multi-Aspect Masker (*MAM*) model architecture to predict and explain A aspect ratings.

weights of the encoder for all E_{a_i}. Finally, the *Classifier* block contains for each aspect a_i a two-layer feedforward neural network followed by a softmax layer to predict the sentiment $\hat{y}_{a_i}$.

3.1 Enabling Interpretability of Masks

The first objective to optimize is the sentiment loss, represented with the cross-entropy between the true aspect sentiment label y_{a_i} and the prediction $\hat{y}_{a_i}$:

$$\ell_{sent} = \sum_{i=1}^{A} \ell_{cross\_entropy}(y_{a_i}, \hat{y}_{a_i}) \qquad (1)$$

However, training the Multi-Aspect Masker to optimize ℓ_{sent} will lead to meaningless sub-masks M_{a_i} because the model tends to focus on certain key-words. Consequently, we guide the model to produce long and meaningful sequences of words, as shown in Figure 1. We propose two regularizers: (1) one to control the number of selected words, and (2) another to encourage consecutive words to belong to the same aspect. For the first term ℓ_{sel}, we calculate the probability p_{sel} of tagging a word as aspect and then compute the cross-entropy with a parameter λ_p. The hyper-parameter λ_p can be interpreted as the prior on the number of selected words among all aspects, which corresponds to the expectation of Binomial(p_{sel}). The optimizer will minimize the difference between p_{sel} and λ_p.

$$p_{sel} = \frac{1}{L} \sum_{\ell=1}^{L} \left(1 - P(m^{\ell}_{a_0} | x^{\ell})\right) \qquad (2)$$
$$\ell_{sel} = \ell_{binary\_cross\_entropy}(p_{sel}, \lambda_p)$$

The second regularizer discourages aspect transition of two consecutive words, by minimizing the mean variation of their aspect distributions M^{ℓ} and $M^{\ell-1}$. We generalize the formulation in Lei et al. (2016) from a hard binary single-aspect selection to a soft probabilistic multi-aspect selection.[4]

$$p_{dis} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{\left|\left|M^{\ell} - M^{\ell-1}\right|\right|_1}{A+1} \qquad (3)$$
$$\ell_{cont} = \ell_{binary\_cross\_entropy}(p_{dis}, 0)$$

We train our Multi-Aspect Masker in an end-to-end manner and optimize the loss $\ell_{MAM} = \ell_{sent} + \lambda_{sel} \cdot \ell_{sel} + \lambda_{cont} \cdot \ell_{cont}$, where λ_{sel} and λ_{cont} control the impact of each constraint.

4 Experiments

In this section, we assess our model in two dimensions: the quality of the explanations, obtained from the induced masks, and the predictive performance. We first measure the quality of the induced sub-masks using human aspect sentence-level annotations, and an automatic topic model evaluation method. In the second experiments, we evaluate our Multi-Aspect Masker (*MAM*) on the multi-aspect sentiment classification task in two different domains.

4.1 Experimental Details

For each model, the review encoder was either a bi-directional single-layer forward recurrent neural network using Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) with 64 hidden units or the multi-channel text convolutional neural network, similar to (Kim et al., 2015), with 3, 5, 7 width filters and 50 feature maps per filter. Each aspect classifier is a two-layer feedforward neural network with ReLU activation function (Nair and Hinton, 2010). We used the 200-dimensional pre-trained word embeddings of Lei et al. (2016) for

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

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 dropout (Srivastava et al., 2014) of 0.1, clipped the gradient norm at 1.0, added L2-norm regularizer with a regularization factor of 10^{-6}, and trained using early stopping. We used Adam (Kingma and Ba, 2015) for training with a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The temperature τ for Gumbel-Softmax distributions was fixed at 0.8. The two regularizer terms and the prior of our model are $\lambda_{sel} = 0.03$, $\lambda_{cont} = 0.03$, 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* dataset. We ran all experiments for a maximum of 50 epochs with a batch-size of 256 on a Titan X GPU.

4.2 Datasets

McAuley et al. (2012) provided 1.5 million 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 models across domains, we crawled 140k hotel reviews from TripAdvisor. Each review contains a five-star rating for each aspect: *Service*, *Cleanliness*, *Value*, *Location*, and *Room*. Compared to beer reviews, hotel reviews are longer, noisier, and less structured, as shown in Appendix A.4 and A.2. Additionally, both datasets do not contain annotations or masks.

As in Bao et al. (2018), we binarized the problem: ratings at three and above are labeled as positive and the rest as negative. We split the datasets into 80/10/10 for train, validation, and test sets. More details about the datasets are in the Appendix.

4.3 Baselines

We compared our Multi-Aspect Masker (*MAM*) with various baselines. We group them in three levels of interpretability:

- ***None***: we cannot identify what parts of the review are important for the prediction;
- ***Coarse-grained***: we observe what parts of the reviews discriminate **all** aspect sentiments, without knowing what part corresponds to what aspect;
- ***Fine-grained***: we identify what parts are used to predict and explain each sentiment aspect.

We first used a simple baseline, *SENT*, that reports the majority sentiment across aspects. Because this information is not available at testing, we trained a model to predict the majority sentiment of a review using Wang and Manning (2012). The second baseline we used is a shared encoder followed by A classifiers that we denote *BASE*. These models do not offer *any interpretability*. We extended it with a shared attention mechanism (Bahdanau et al., 2015) after the encoder, noted *SAA*, 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*: to visualize what sequences of words a model focuses on and to predict the aspect sentiments. To this end, we included other baselines: two trained *separately* for each aspect and two trained with a *multi-aspect* sentiment loss. We employed for the first ones the well-known *NB-SVM* of Wang and Manning (2012) for sentiment analysis tasks, and the Single-Aspect Masker (*SAM*) model from Lei et al. (2016), each trained *separately* for each aspect.

The two last methods are composed of a separate encoder, attention mechanism, and classifier for each aspect. We utilized two types of attention mechanisms: additive (Bahdanau et al., 2015) and sparse (Martins and Astudillo, 2016). We call them Multi-Aspect Attentions (MAA) and Sparse-Attentions (MASA), respectively. Diagrams of the baselines can be found in Appendix A.3.

Finally, to demonstrate that the induced sub-masks $M_{a_1}, ..., M_{a_A}$ computed from *MAM* (1) bring fine-grained interpretability, and (2) are meaningful for other models to improve final predictions, we extracted and concatenated the masks to the word embeddings, resulting in contextualized embeddings (Peters et al., 2018). We trained *BASE* with the contextualized embeddings and denote this variant *MAM*^{C}. Its advantage is to be smaller and has faster inference than *MAM*.

4.4 Mask Interpretability

In this section, we first verify whether inferred masks $M_{a_1}, M_{a_2}, ..., M_{a_A}$ of *MAM* are meaningful and interpretable, compared to other models.

4.4.1 Mask Precision

Evaluating explanations that have short and coherent pieces of text is challenging because there is no gold standard for reviews. McAuley et al. (2012) provided 994 beer reviews with aspect sentence-

Model	Precision / % Highlighted words		
	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%
MAM	**96.6** / 7%	**81.7** / 7%	**96.7** / 14%

* Model trained separately for each aspect.

Table 1: Performance on human evaluation. Precision of selected words for each aspect for the *Beer* dataset. Percentage of words indicates the number of highlighted words of the full review, as in Lei et al. (2016).

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

* Model trained separately for each aspect.

† Metric that correlates best with human judgment (Lau and Baldwin, 2016).

Table 2: Performance on automatic evaluation. Average Topic Coherence (NPMI) across different top-N words for each dataset. We consider each aspect a_i a topic and use the masks/attentions to compute $P(w|a_i)$.

level annotations, although our model computes masks at a finer level. Each sentence of the dataset is annotated with one aspect label, indicating what aspect that sentence covers. We evaluate the precision of words highlighted by each model, as in Lei et al. (2016). We used trained models on the *Beer* dataset and extracted a similar number of selected words for a fair comparison. We report the results of the models in Lei et al. (2016): *NB-SVM*, the Single-Aspect Attention (*SAA*), and Single-Aspect Masker (*SAM*) – trained *separately* for each aspect because they find *hard* masks for a *single* aspect.

Table 1 presents the precision of the masks and attentions computed on sentence-level aspect annotations. We show that the generated sub-masks obtained with our Multi-Aspect Masker (*MAM*) correlate best with human judgment. In comparison to *SAM*, the *MAM* model obtains significantly higher precision with an average of $+1.13$. Interestingly, *NB-SVM* and attention models (*SAA*, *MASA*, *MAA*) perform poorly compared with mask models: especially *MASA*, which focuses only on a couple of words due to the sparseness of the attention.

4.4.2 Mask Semantic Coherence

In addition to evaluating masks with human annotations, we computed their semantic interpretability. According to Aletras and Stevenson (2013);Lau et al. (2014), NPMI is a good metric for 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 computed the topic coherence over several cardinalities N, and report the results and the average;[5] the authors claim that the mean leads to a more stable and robust evaluation.

The results are shown in Table 2. We show that computed masks by *MAM* obtain the highest mean NPMI and, on average, 20% superior results in both datasets, while only needing a single training. Our model *MAM* significantly outperforms *SAM* and attention models (*MASA* and *MAA*) for $N \geq 20$ and $N = 5$. For $N = 10$ and $N = 15$, *MAM* obtains higher scores in two out of four cases ($+.033$ and $+.009$), and for the two others, the difference is 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 finds 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.

4.5 Multi-Aspect Sentiment Classification

In the previous section, we showed that the inferred masks of *MAM* are significantly more accurate and semantically coherent than other models. In the next experiments, we inquire whether masks can become a benefit rather than a cost in performance for multi-aspect sentiment classification.

[5] We include N=30 but do not show it in the table due to space limit. Appendix A.1 details the derivation of the topics.

Interp.		Model	Params	F1 Score				
				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 et al. (2016)	$4 \cdot 644k$	76.62	72.93	77.94	75.70	79.91
	MASA	Emb_{200} + Enc_{LSTM} + $\text{A}^{\text{Sparse}}_{\text{Aspect-wise}}$ + Clf	$611k$	77.62	72.75	79.62	75.81	82.28
	MAA	Emb_{200} + Enc_{LSTM} + $\text{A}_{\text{Aspect-wise}}$ + Clf	$611k$	78.50	74.58	79.84	77.06	82.53
	MAM	Emb_{200} + Masker + Enc_{CNN} + Clf (Ours)	$289k$	78.55	74.87	79.93	77.39	82.02
	MAM$^{\text{C}}$	**Emb_{200+4} + Enc_{CNN} + Clf (Ours)**	$191k$	**78.94**	**75.02**	**80.17**	**77.86**	**82.71**

Table 3: Performance of the multi-aspect sentiment classification task for the ***Beer*** dataset.

4.5.1 Beer Reviews

We report the macro F1 score and individual one for each aspect A_i. Table 3 presents the results for the *Beer* dataset. The Multi-Aspect Masker (*MAM*) performs better on average than all baselines and provides fine-grained interpretability. Moreover, *MAM* has two times fewer parameters than aspect-wise attention models (*MAA* and *MASA*).

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

NB-SVM, which offers fine-grained interpretability and is trained *separately* for each aspect, significantly underperform compared to *BASE* and, surprisingly, to *SENT*. To understand why the majority sentiment baseline performs better, we calculated the sentiment correlation between any pair of aspects of the *Beer* dataset. We found an average correlation of 71.8%. Therefore, by predicting the sentiment of one aspect correctly, it is likely that other aspects share the same polarity. We suspect that the linear model *NB-SVM* cannot capture the correlated relationships between aspects, unlike non-linear (neural) models, having a higher capacity.

Shared attention models (*SAA*) perform better than *BASE* but provide only coarse-grained interpretability. *SAM* is outperformed by all other models besides *SENT*, *BASE*, and *NB-SVM*.

4.5.2 Model Robustness - Hotel Reviews

We check the robustness of our model on another domain. Table 4 presents the results of the *Hotel* dataset. The contextualized variant *MAM^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 *MAM* is again meaningful because it increases the performance and adds interpretability to *BASE*.

Regarding *MAM*, we see that it performs slightly worse than aspect-wise attention models *MASA* and *MAA* but has 2.5 times fewer parameters.

A visualization of a hotel review with the masks $M_{a_1}, ..., M_{a_A}$ and attentions is available in Figure 1. Not only do 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, with the masks and attentions computed by different models, can be found in Appendix A.4.

To summarize, we showed that the two regularizers in *MAM* guide the model to produce high-quality masks as explanations while performing comparably to strong attention models in terms of prediction performance. However, we demonstrate that including the inferred masks into word embeddings and train a simpler baseline model, achieves the best performance across two datasets and brings at the same time fine-grained interpretability.

4.5.3 Single versus Multi -Aspect Masker

SAM is the neural model obtaining the lowest relative macro F1 score in the two datasets compared with *MAM^C*: a difference of -2.32 and -3.27 for the *Beer* and *Hotel* datasets, respectively. As in Section 4.5.1, we compute the average correlation between any pair of aspects and found a correlation of 71.8% and 63.0%. Therefore, there is a high correlation between the aspect ratings in the same

Interp.		Model	Params	F1 Score 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	$\text{Emb}_{300} + \text{Enc}_{\text{CNN}} + \text{Clf}$	$263k$	90.30	92.91	93.55	94.12	76.65	94.29
Coarse-grained	SAA	$\text{Emb}_{300} + \text{Enc}_{\text{CNN}} + \text{A}_{\text{Shared}} + \text{Clf}$	$301k$	90.12	92.73	93.55	93.76	76.40	94.17
		$\text{Emb}_{300} + \text{Enc}_{\text{LSTM}} + \text{A}_{\text{Shared}} + \text{Clf}$	$270k$	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 et al. (2016)	$5 \cdot 824k$	87.52	91.48	91.45	92.04	70.80	91.85
	MASA	$\text{Emb}_{200} + \text{Enc}_{\text{LSTM}} + \text{A}^{\text{Sparse}}_{\text{Aspect-wise}} + \text{Clf}$	$1010k$	90.23	93.11	93.32	93.58	77.21	93.92
	MAA	$\text{Emb}_{300} + \text{Enc}_{\text{LSTM}} + \text{A}_{\text{Aspect-wise}} + \text{Clf}$	$1010k$	90.21	92.84	93.34	93.78	76.87	94.21
	MAM	$\text{Emb}_{300} + \text{Masker} + \text{Enc}_{\text{CNN}} + \text{Clf}$ (Ours)	$404k$	89.94	92.84	92.95	93.91	76.27	93.71
	MAM$^{\text{C}}$	**Emb_{300+5} + Enc_{CNN} + Clf (Ours)**	$267k$	**90.79**	**93.38**	**93.82**	**94.55**	**77.47**	**94.71**

Table 4: Performance of the multi-aspect sentiment classification task for the ***Hotel*** dataset.

review, making it challenging to learn their justifications directly. Following the observations of Lei et al. (2016), this highlights that *SAM* suffers from high correlation, unlike *MAM*.

Therefore, we compare our proposed model in a setting where aspect ratings are less correlated, although it does not reflect the real distribution of aspect ratings. We employ the *decorrelated* subset of the *Beer* reviews from Lei et al. (2016); Li et al. (2016). It has an average correlation of 27.2% and the aspect *Taste* removed. The results and a comparison of the datasets are shown in Appendix A.2.

We find similar trends but stronger results: *MAM* significantly generates better masks and achieves higher F1 scores than *SAM* and attention models. The contextualized variant *MAM$^{\text{C}}$* further improves the performance. Figure 3 shows an example of generated masks and attentions. *MAM* highlights concrete information related to aspects, whereas *SAM* misses some important words.

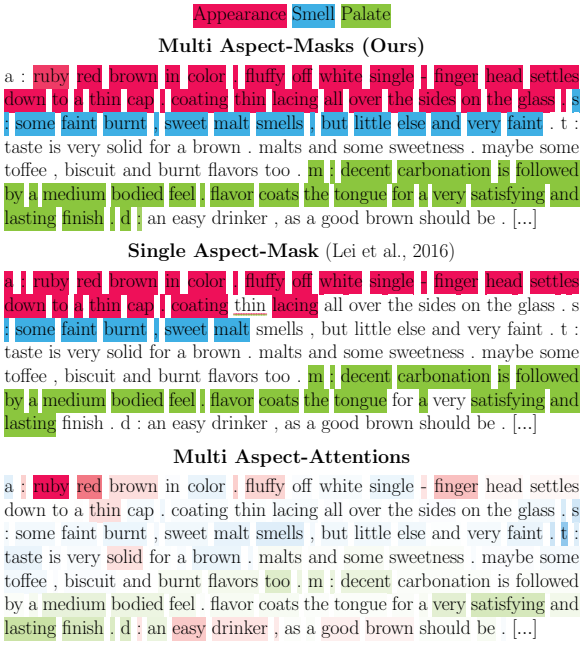

Figure 3: A beer review with masks and attentions from models trained on the *decorrelated* dataset. *MAM* selects all the words corresponding to the aspects. *SAM* only highlights the most crucial information. *MAA* noisily labels words, most with low confidence.

5 Conclusion

Providing explanations along automated predictions can carry a lot more impact, increase transparency, and be even necessary for some fields. Past work has proposed to generate explanations from reviews to support users' ratings, but often the resulting justifications are noisy and might rely on unrealistic assumptions on the data. We proposed Multi-Aspect Masker (MAM), a multi-task learning neural model that jointly predicts multi-aspect ratings from reviews with textual explanations, in the absence of any explicit annotations. According to human annotations and automatic evaluation, inferred masks are more accurate and semantically coherent than those produced by attention and strong baselines.

In earlier work, interpretability of explanations came at the expense of a loss in accuracy. We have shown that MAM even improves accuracy over the best baselines. It is the first technique that delivers both the best explanations and highest accuracy.

As future work, we plan to use masks for recommendation: to generate personalized and relevant explanations to support recommendations and build better user and item profiles. Further, masks embed opinions in reviews, and thus could be used to produce structured summaries across multiple aspects.

References

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

Diego Antognini and Boi Faltings. 2020. Hotelrec: a novel very large-scale hotel recommendation dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4917–4923, Marseille, France. European Language Resources Association.

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.

G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gunter Narr Verlag.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Aurélie Herbelot and Eva Maria Vecchi. 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*.

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

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.

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.

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*.

Jey Han Lau and Timothy Baldwin. 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*, pages 483–487.

Jey Han Lau, David Newman, and Timothy Baldwin. 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*.

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.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

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*.

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

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

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

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing*, pages 455–466.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2227–2237, New Orleans.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 90–94, Jeju Island, Korea.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2054.

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 $\boldsymbol{w}_{\boldsymbol{a_i}}^{\boldsymbol{N}}$, the Normalized Pointwise Mutual Information (Bouma, 2009) coherence score is:

$$\text{NPMI}(\boldsymbol{w}_{\boldsymbol{a_i}}^{\boldsymbol{N}}) = \sum_{j=2}^{N} \sum_{k=1}^{j-1} \frac{\log \frac{P(w_{a_i}^k, w_{a_i}^j)}{P(w_{a_i}^k) P(w_{a_i}^j)}}{-\log P(w_{a_i}^k, w_{a_i}^j)} \quad (4)$$

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 et al. (2012): we first computed the averages across all aspect words for each word w: $b_w = \frac{1}{|A|} \sum_{i=1}^{|A|} P(w|a_i)$. 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: $\hat{P}(w|a_i) = P(w|a_i) - b_w$.

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 4.5.3. Table 7 presents descriptive statistics of *Beer* and *Hotel* datasets with the *decorrelated* subset of beer reviews from Lei et al. (2016); Li et al. (2016). The results of the multi-aspect sentiment classification experiment are shown in Table 8. Table 9 contains the results of the intruder tasks (see Appendix A.1).

A.3 Baseline Architectures

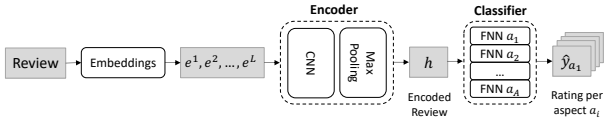

Figure 4: Baseline model Emb + Enc_{CNN} + Clf (*BASE*).

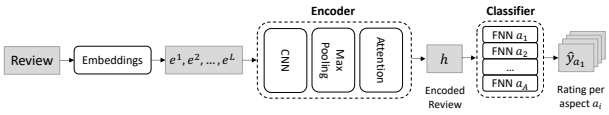

Figure 5: Baseline model Emb + Enc_{CNN} + A_{Shared} + Clf (*SAA*, CNN variant).

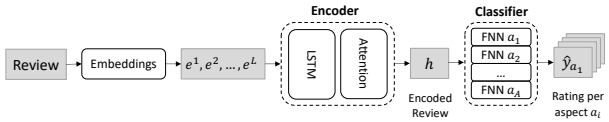

Figure 6: Baseline model Emb + Enc_{LSTM} + A_{Shared} + Clf (*SAA*, LSTM variant).

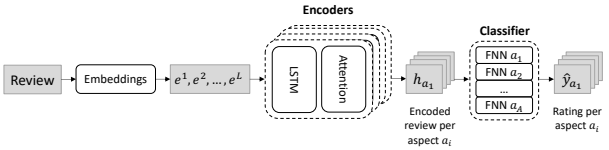

Figure 7: Baselines Emb + Enc_{LSTM} + $\text{A}^{\text{[Sparse]}}_{\text{Aspect-wise}}$ + Clf. Attention is either additive (*MAA*) or sparse (*MASA*).

A.4 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-Aspect Masker (*MAM*), the Single-Aspect Masker (*SAM*) (Lei et al., 2016), and two attention models with additive and sparse attention, called Multi-Aspect Attentions (*MAA*) and Multi-Aspect Sparse-Attentions (*MASA*), respectively (more details in 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 8, Figure 9, Figure 10, and Figure 11, respectively.

	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
	MAM (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
	MAM (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**
	MAM (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
	MAM (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, *MAM* 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
	MAM (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
	MAM (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**
	MAM (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
	MAM (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
	MAM (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 *MAM* do not contain any intruder.

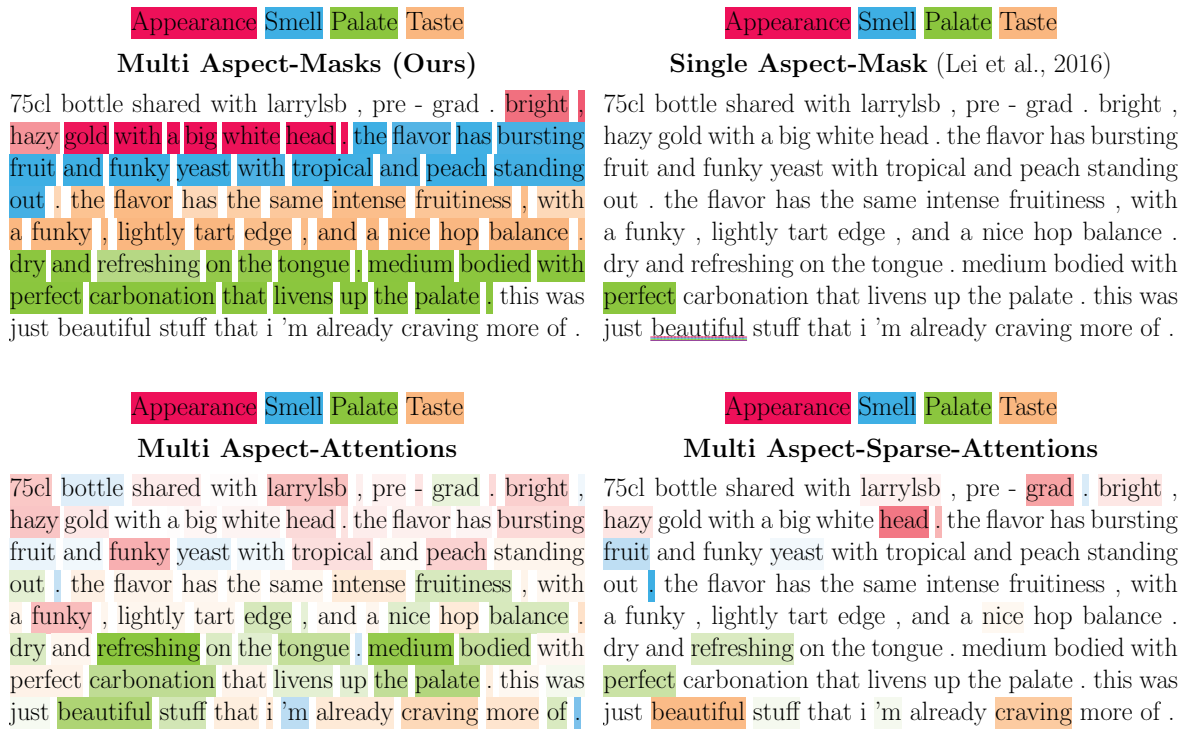

Figure 8: A sample review from the *Beer* dataset, with computed masks from different methods. *MAM* achieves near-perfect annotations, while *SAM* highlights only two words where one is ambiguous with respect to the four aspects. *MAA* mixes between the aspect *Appearance* and *Smell*. *MASA* identifies some words but lacks coverage.

Multi Aspect-Masks (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 !

Single Aspect-Mask (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 !

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 !

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 9: *MAM* 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.

Figure 10: *MAM* 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*.

Service Cleanliness Value Location Room

Multi-Aspect 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-Mask (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 11: Our *MAM* 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.

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	Emb_{200} + Enc_{CNN} + Clf	$173k$	78.23	78.38	80.86	75.47
Coarse-grained	SAA	Emb_{200} + Enc_{CNN} + A_{Shared} + Clf	$196k$	78.19	77.43	80.96	76.16
		Emb_{200} + Enc_{LSTM} + A_{Shared} + 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 et al. (2016)	$3 \cdot 644k$	77.06	77.36	78.99	74.83
	MASA	Emb_{200} + Enc_{LSTM} + $\text{A}^{\text{Sparse}}_{\text{Aspect-wise}}$ + Clf	$458k$	78.82	77.35	81.65	77.47
	MAA	Emb_{200} + Enc_{LSTM} + $\text{A}_{\text{Aspect-wise}}$ + Clf	$458k$	78.96	78.54	81.56	76.79
	MAM	Emb_{200} + Masker + Enc_{CNN} + Clf (Ours)	$274k$	79.32	78.58	81.71	77.66
	MAM$^{\text{C}}$	**Emb_{200+4} + Enc_{CNN} + 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
	MAM (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
	MAM (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
	MAM (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, MAM has only one word considered as an intruder, followed by MASA with SAM (two) and MAA (six).