ABSTRACT

Neural models achieved considerable improvement for many natural language processing tasks, but they offer little transparency, and interpretability comes at a cost. In some domains, automated predictions without justifications have limited applicability. Recently, progress has been made regarding single-aspect sentiment analysis for reviews, where the ambiguity of a justification is minimal. In this context, a justification, or mask, consists of (long) word sequences from the input text, which suffice to make the prediction. Existing models cannot handle more than one aspect in one training and induce binary masks that might be ambiguous. In our work, we propose a neural model for predicting multi-aspect sentiments for reviews and generates a probabilistic multi-dimensional mask (one per aspect) simultaneously, in an unsupervised and multi-task learning manner. Our evaluation shows that on three datasets, in the beer and hotel domain, our model outperforms strong baselines and generates masks that are: strong feature predictors, meaningful, and interpretable.

1 INTRODUCTION

Neural networks 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 & Wallace, 2019].

In many domains, automated predictions have a real impact on the final decision, such as treatment options in the field of medicine. Therefore, it is important to provide the underlying reasons for such a decision. We claim that integrating interpretability in a (neural) model should supply the reason of the prediction and should yield better performance. However, justifying a prediction might be ambiguous and challenging. Prior work includes various methods that find the justification in an input text — also called rationale or mask of a target variable. The mask is defined as one or multiple pieces of text fragments from the input text. Each should contain words that altogether are short, coherent, and alone sufficient for the prediction as a substitute of the input [Lei et al., 2016].

Many works have been applied to single-aspect sentiment analysis for reviews, where the ambiguity of a justification is minimal. In this case, we define an aspect as an attribute of a product or service (Giannakopoulos et al., 2017), such as Location or Cleanliness for the hotel domain. There are three different methods to generate masks: using reinforcement learning with a trained model [Li et al., 2016], generating rationales in an unsupervised manner and jointly with the objective function [Lei et al., 2016], or including annotations during training [Bao et al., 2018; Zhang et al., 2016].

In the rest of the paper, we will use the terms mask, justification and rationale interchangeably.
However, these models generate justifications that are 1) only tailored for one aspect, and 2) expressed as a hard (binary) selection of words. A review text reflects opinions about multiple topics a user cares about. It appears reasonable to analyze multiple aspects with a multi-task learning setting, but a model must be trained as many times as the number of aspects. A hard 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.", the word "room" refers to the aspects Room, Cleanliness and Location. Finally, collecting human-provided rationales at scale is expensive and thus impractical.

In this work, we study interpretable multi-aspect sentiment classification. We describe an architecture for predicting the sentiment of multiple aspects while generating a probabilistic (soft) multi-dimensional mask (one dimension per aspect) jointly, in an unsupervised and multi-task learning manner. We show that the induced mask is beneficial for identifying simultaneously what parts of the review relate to what aspect, and 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 a cost of reduced accuracy. In contrast, our evaluation shows that on three datasets, in the beer and hotel domain, our model outperforms strong baselines and generates masks that are: strong feature predictors, meaningful, and interpretable compared to attention-based methods and a single-aspect masker. 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 anymore.

The contributions of this work can be summarized as follow:

- We propose a Multi-Aspect Masker (MAM), an end-to-end neural model for multi-aspect sentiment classification that provides fine-grained interpretability in the same training. Given a text review as input, the model generates a probabilistic multi-dimensional mask, with one dimension per aspect. It predicts the sentiments of multiple aspects, and highlights long sequences of words justifying the current rating prediction for each aspect;
- We show that interpretability does not come at a cost: our final model significantly outperforms strong baselines and attention models, both in terms of performance and mask coherence. Furthermore, the level of interpretability is controllable using two regularizers;
- Finally, we release a new dataset for multi-aspect sentiment classification, which contains 140k reviews from TripAdvisor with five aspects, each with its corresponding rating.

2 RELATED WORK

2.1 INTERPRETABILITY

Developing interpretable models is of considerable interest to the broader research community, even more pronounced with neural models. Many works analyzed and visualized state activation [Karpathy et al., 2015; Li et al., 2016a; Montavon et al., 2018], learned sparse and interpretable word vectors [Faruqui et al., 2015a; Herbelot & Vecchi, 2015] or analyzed attention [Clark et al., 2019; Jain & Wallace, 2019]. Our work differs from these 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 prediction.

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 et al., 2015] over hard [Luong et al., 2015] and sparse ones [Martins & Astudillo, 2016]. According to Jain & Wallace (2019); Serrano & Smith (2019), standard attention modules noisily predict input importance; the weights cannot provide safe and meaningful explanations.

---

2We will make the code and data available.
Our model differs in two ways from attention mechanisms: our loss includes two regularizers to favor long word sequences for interpretability; the normalization is not done over the sequence length.

2.3 Multi-Aspect Sentiment Classification

Review multi-aspect sentiment classification is sometimes seen as a sub-problem (Wang et al., 2010; McAuley et al., 2012; Pappas & Popescu-Belis, 2014), by utilizing heuristic-based methods or topic models. Recently, neural models achieved 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 overall rating, aspect, and user embeddings. The disadvantage of these methods is their limited interpretability, as they rely on many features in addition to the review text.

2.4 Rationale-Based Models

The idea of including human rationales during training has been explored in (Zhang et al., 2016; Marshall et al., 2015; 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. (2016b) 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. (2016b) developed a post-training method that removes words from a review text until another trained model changes its prediction. Lei et al. (2016) provides a model that learns an aspect sentiment and its rationale jointly, but 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

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, is composed of three components: 1) a \textit{Masker} module that computes a probability distribution over aspects for each word, resulting in \( A + 1 \) different masks (including one for \textit{not-aspect}); 2) an \textit{Encoder} that learns a representation of a review conditioned on the induced masks; 3) a \textit{Classifier} that predicts the target variables. The overall model architecture is shown in Figure 1. Our framework generalizes for other tasks, and each neural module is interchangeable with other models.

The \textit{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 neural networks. In our case, we chose a convolutional network because it led to a smaller model, faster training, and empirically, performed similarly to recurrent
models. Let \( a_i \) denote the \( i^{th} \) aspect for \( i = 1, \ldots, A \), 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 \( x^\ell \) as:

\[
P(M|X) = \prod_{\ell=1}^{L} P(M^\ell|x^\ell) = \prod_{\ell=1}^{L} \prod_{i=0}^{A} P(m^\ell_{ai}|x^\ell)
\]

(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 \( m^\ell_{ai} \) using the Straight Through Gumbel Softmax \( \text{Jang et al. (2017)} \) \( \text{Maddison et al. (2017)} \), to approximate sampling from a categorical distribution. We model the parameters of each Gumbel Softmax distribution \( M^\ell \) with a single-layer feedforward neural network followed by applying a log softmax, which induces the log-probabilities of the \( \ell^{th} \) distribution: \( \omega_k = \log(\text{softmax}(W x^\ell + b)) \). \( W \) and \( b \) are shared across all tokens, to have a constant number of parameters with respect to the sequence length. We control the sharpness of the distributions with the temperature parameter \( \tau \). Compared to attention mechanisms, the word importance is a probability distribution over the targets: \( \sum_{\ell=0}^{T} P(m^\ell_{ai}|x^\ell) = 1 \), instead of a normalization over the sequence length, \( \sum_{\ell=1}^{L} P(a^\ell|x^\ell) = 1 \).

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

\[
M_{ai} = P(m^1_{ai}|x^1), P(m^2_{ai}|x^2), \ldots, P(m^L_{ai}|x^L)
\]

(2)

We weight the word embeddings by their importance towards an aspect \( a_i \) with the induced sub-masks, such that \( E_{ai} = M_{ai} \odot E = P(m^1_{ai}|x^1) \cdot e_1, P(m^2_{ai}|x^2) \cdot e_2, \ldots, P(m^L_{ai}|x^L) \cdot e_L \). Thereafter, each modified embedding \( E_{ai} \) 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.

The Encoder module includes a convolutional neural network, for the same reasons as earlier, followed by a max-over-time pooling to obtain a fixed-length feature vector. It produces the hidden representation \( h_{ai} \) for each aspect \( a_i \). To exploit commonalities and differences among aspects, we share the weights of the encoders for all \( E_{ai} \). Finally, the Classifier block contains for each aspect \( a_i \) a two-layer feedforward neural networks followed by a softmax layer to predict the sentiment \( \hat{y}_{ai} \).

3.1 Interpretable Masks

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

\[
\ell_{\text{sent}} = \sum_{i=1}^{A} \ell_{\text{cross-entropy}}(y_{ai}, \hat{y}_{ai})
\]

(3)

Training Multi-Aspect Masker to optimize \( \ell_{\text{sent}} \) will lead to meaningless sub-masks \( M_{ai} \), as 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[2]. We propose two regularizers: the first controls the number of selected words, and the second favors consecutive words belonging to the same aspect. For the first term \( \ell_{\text{sel}} \), we calculate the probability \( p_{\text{sel}} \) of tagging a word as aspect and then compute the cross-entropy with a parameter \( \lambda_p \). The hyper-parameter \( \lambda_p \) can be interpreted as the prior on the number of selected words among all aspects, which corresponds to the expectation of Binomial(\( p_{\text{sel}} \)), as the optimizer will try to minimize the difference between \( p_{\text{sel}} \) and \( \lambda_p \).

\[
p_{\text{sel}} = \frac{1}{L} \sum_{\ell=1}^{L} \left( 1 - P(m^\ell_{a_0}|x^\ell) \right)
\]

\[
\ell_{\text{sel}} = \ell_{\text{binary-cross-entropy}}(p_{\text{sel}}, \lambda_p)
\]

The second regularizer discourages aspect transition between two consecutive words, by minimizing the mean variation of two consecutive aspect distributions. We generalize the formulation in \( \text{Lei et al. (2016)} \), from a hard binary single-aspect selection, to a soft probabilistic multi-aspect selection.

\[
p_{\text{dis}} = \frac{1}{L} \sum_{\ell=1}^{L} \left[ \frac{1}{A+1} \sum_{a=0}^{A} | P(m^\ell_{ai}|x^\ell) - P(m^{\ell-1}_{ai}|x^{\ell-1}) | \right]
\]

\[
\ell_{\text{cont}} = \ell_{\text{binary-cross-entropy}}(p_{\text{dis}}, 0)
\]

3: if \( P(m^\ell_{a_0}|x^\ell) \approx 1.0 \), it implies that \( \sum_{i=1}^{A} P(m^\ell_{ai}|x^\ell) \approx 0 \) and consequently, \( e_{ai} \approx 0 \).
Under review as a conference paper at ICLR 2020

| Attention model | Multi-Aspect Masker |
|-----------------|---------------------|
| Trained on $\ell_{sent}$ and no constraint | Trained on $\ell_{sent}$ with $\lambda_p$, $\ell_{sel}$, and $\ell_{cont}$ |

| 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 ushiana , 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 | 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 ushiana , 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 |

Aspect Changes: 30

Aspect Changes: 5

Figure 2: Justifications obtained for a hotel review, with an attention model and Multi-Aspect Masker, where the colors represent the aspects: **Service, Cleanliness, Value, Location, and Room.** Masks lead to mostly long sequences describing clearly each aspect (one switch per aspect), while attention to many short and interleaving sequences (30 changes between aspects), where most relate to noise or multiple aspects. Highlighted words correspond to the highest aspect-attention scores above $1/L$ (i.e., more important than a uniform distribution), and the aspect $a_t$ maximizing $P(m^*_a|x^f)$.

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

4 Experiments

In this section, we assess our model on two dimensions: the predictive performance and the quality of the induced mask. We first evaluate Multi-Aspect Masker on the multi-aspect sentiment classification task. In a second experiment, we measure the quality of the induced sub-masks using aspect sentence-level annotations, and an automatic topic model evaluation method.

4.1 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. [Lei et al., 2016] modified the dataset to suit the requirements of their method. As a consequence, the obtained subset, composed of 280k reviews, does not reflect the real data distribution: it contains only the first three aspects, and the sentiment correlation between any pair of aspects is decreased significantly (27.2% on average, instead of 71.8% originally). We denote this version as the Filtered Beer dataset, and the original one as the Full Beer dataset.

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. The average correlation between aspects is high (63.0% on average). Compared to beer reviews, hotel reviews are longer, noiser, and less structured, as shown in Appendix A.3.

As in [Bao et al., 2018], we binarize the problem: ratings at three and above are labeled as positive and the rest as negative. We further divide the datasets into 80/10/10 for train, development, and test subsets (more details in Appendix A.1).

The detailed experimental setup is described in Appendix A.2.

For the three first aspects, they trained a simple linear regression model to predict the rating of an aspect given the others and then selected reviews with the largest prediction error.
To demonstrate that the induced sub-masks \( M_i \) improve final predictions, and 2) bring fine-grained interpretability, we extracted and concatenated significant improvement of at least provides fine-grained interpretability. For the synthetic Filtered Beer wise attention models, performs better on average than all other baselines on both datasets, and

We find that our Multi-Aspect Masker model, with in Table 1 and Table 2. We do not report SVM results for the latter, due to the lengthy training time. (where there are assumptions on the data distribution) and the real-world A

4.2 BASELINES

We compared our Multi-Aspect Masker (MAM) with various baselines. The first model is a shared encoder followed by \( A \) classifiers, that we denote \( \text{Emb} + \text{Enc} + \text{Clf} \). This model does not offer any interpretability. We extended it with a shared attention mechanism [Bahdanau et al.] after the encoder, noted \( A_{\text{Shared}} \), that provides a coarse-grained interpretability: for all aspects, the network 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 sentiment predictions. 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: a bigram SVM combined with tf-idf, and the Single Aspect-Mask (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 mechanism: additive [Bahdanau et al., 2015], and sparse [Martins & Astudillo, 2016]. We call each variant Multi Aspect-Attentions (MAA) and Multi Aspect-Sparse-Attentions (MASA). Diagrams for the baselines can be found in Appendix A.5.

4.3 MULTI-ASPECT SENTIMENT CLASSIFICATION

In this section, we enquire whether fine-grained interpretability can become a benefit rather than a cost. We group the models and baselines in three different levels of interpretability:

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

4.3.1 BEER REVIEWS

Overall F1 scores (macro and for each aspect \( A_i \)) for the controlled-environment Filtered Beer (where there are assumptions on the data distribution) and the real-world Full Beer dataset are shown in Table 1 and Table 2. We do not report SVM results for the latter, due to the lengthy training time.

We find that our Multi-Aspect Masker model, with 1.7 to 2.1 times fewer parameters than aspect-wise attention models, performs better on average than all other baselines on both datasets, and provides fine-grained interpretability. For the synthetic Filtered Beer dataset, MAM achieves a significant improvement of at least 0.36 macro F1 score, and 0.05 for the Full Beer one.

To demonstrate that the induced sub-masks \( M_{a_1}, \ldots, M_{a_d} \) are 1) meaningful for other models to improve final predictions, and 2) bring fine-grained interpretability, we extracted and concatenated

| Interp. | Model | Params | Macro | \( A_1 \) | \( A_2 \) | \( A_3 \) |
|---------|-------|--------|-------|--------|--------|--------|
| None    | \( \text{Emb}^200 + \text{Enc}_{\text{CNN}} + \text{Clf} \) | 173k | 78.23 | 78.38 | 80.86 | 75.47 |
| Coarse-grained | \( \text{Emb}^200 + \text{Enc}_{\text{CNN}} + A_{\text{Shared}} + \text{Clf} \) | 196k | 78.19 | 77.43 | 80.96 | 76.16 |
|          | \( \text{Emb}^200 + \text{Enc}_{\text{LSTM}} + A_{\text{Shared}} + \text{Clf} \) | 186k | 78.16 | 75.88 | 81.25 | 77.36 |
|          | SVM   |        | 3 \cdot 9M | 67.94 | 72.56 | 65.78 | 65.47 |
|          | SAM (Lei et al., 2016) | 3 \cdot 644k | 77.06 | 77.36 | 78.99 | 74.83 |
|          | \( \text{Emb}^200 + \text{Enc}_{\text{LSTM}} + A_{\text{Aspect-wise}} + \text{Clf} \) | 458k | 78.82 | 77.35 | 81.65 | 77.47 |
|          | \( \text{Emb}^200 + \text{Enc}_{\text{LSTM}} + A_{\text{Aspect-wise}} + \text{Clf} \) | 458k | 78.96 | 78.54 | 81.56 | 76.79 |
|          | \( \text{Emb}^200 + \text{Masker} + \text{Enc}_{\text{CNN}} + \text{Clf} (\text{Ours}) \) | 274k | 79.32 | 78.58 | 81.71 | 77.66 |
|          | \( \text{Emb}^200+3 + \text{Enc}_{\text{CNN}} + \text{Clf} (\text{Ours}) \) | 175k | **79.66** | **78.74** | **82.02** | **78.22** |
the masks to the word embeddings, resulting in contextualized embeddings (Peters et al. 2018). We trained a simple Encoder-Classifier (last row) with the contextualized embeddings, which has approximately 1.5 times fewer parameters than MAM. We achieved a macro F1 score absolute improvement of 0.34 compared to MAM, and 1.43 compared to the non-contextualized variant for the Filtered Beer dataset; for the Full Beer one, the performance increases by 0.39 and 2.49 respectively. Similarly, each individual aspect $A_i$ F1 score of MAM is improved to a similar extent.

We provide in Appendix A.3.1 and A.3.2 visualizations of reviews with the computed sub-masks $M_{a_1}, \ldots, M_{a_A}$ and attentions by different models. Not only do sub-masks enable the reach of higher performance; they better capture parts of reviews related to each aspect compared to other methods. Both SVM and SAM (offering fine-grained interpretability and trained separately for each aspect) significantly underperform compared to the Encoder-Classifier. This result is expected: the goal of SAM is to provide rationales at the price of performance (Lei et al. 2016). Shared attention models perform similarly to the Encoder-Classifier, but provide only coarse-grained interpretability.

However, in the Full Beer dataset, SAM obtains better results than the Encoder-Classifier baseline, which is outperformed by all other models. It might be counterintuitive that SAM performs better, but we claim that its behavior comes from the high correlation between aspects: SAM select words that should belong to aspect $a_j$ to predict the sentiment of aspect $a_j$ ($a_j \neq a_i$). Moreover, in Section 4.3.2 we show that a single-aspect mask from SAM cannot be employed for interpretability. These results emphasize the ease of the Filtered Beer dataset compared to the Full Beer one, because of the assumptions not holding in the real data distribution.

### 4.3.2 Model Robustness - Hotel Reviews

On the Hotel dataset, the learned mask $M$ from Multi-Aspect Masker is again meaningful, by increasing the performance and adding interpretability. The Encoder-Classifier with contextualized embeddings (last row) outperforms all other models significantly, with an absolute macro F1 score improvement of 0.49. Moreover, it achieves the best individual F1 score for each aspect $A_1, \ldots, A_5$.

Visualizations of reviews, with masks and attentions, are available in Appendix A.3.3. The interpretability comes from the long sequences that MAM identifies, unlike attention models. In comparison, SAM lacks coverage and suffers from ambiguity due to the high correlation between aspects.

We observe that Multi-Aspect Masker performs slightly worse than aspect-wise attention models, with 2.5 times fewer parameters. We emphasize that using the induced masks in the Encoder-Classifier already achieves the best performance.

The Single Aspect-Mask obtains the lowest relative macro F1 score of the three datasets: a difference of $-3.27$, $-2.6$ and $-2.32$ for the Filtered Beer and Full Beer dataset respectively. This proves that the model is not meant to provide rationales and increase the performance simultaneously.
Table 3: Performance of the best models of each architecture for the Hotel dataset.

| Interp.       | Model                                      | Params | Macro | $A_1$  | $A_2$  | $A_3$  | $A_4$  | $A_5$  |
|---------------|--------------------------------------------|--------|-------|--------|--------|--------|--------|--------|
| None          | Emb$_{300}$ + Enc$_{CNN}$ + Clf           | 263$k$ | 90.30 | 92.91  | 93.55  | 94.12  | 76.65  | 94.29  |
| Coarse-grained| Emb$_{300}$ + Enc$_{CNN}$ + A$_{Shared}$ + Clf | 301$k$ | 90.12 | 92.73  | 93.55  | 93.76  | 76.40  | 94.17  |
|               | Emb$_{300}$ + Enc$_{LSTM}$ + A$_{Shared}$ + Clf | 270$k$ | 88.22 | 91.13  | 91.92  | 93.33  | 71.40  | 93.06  |
| Fine-grained  | SVM                                        | 5 · 9$M$ | 88.12 | 90.40  | 91.35  | 91.58  | 75.66  | 91.63  |
|               | SAM (Lei et al., 2016)                     | 5 · 824$k$ | 87.52 | 91.48  | 91.45  | 92.04  | 70.80  | 91.85  |
|               | Emb$_{200}$ + Enc$_{LSTM}$ + A$_{Aspect-wise}$ + Clf | 1010$k$ | 90.24 | 93.11  | 93.32  | 93.58  | 77.21  | 93.92  |
|               | Emb$_{300}$ + Masker + Enc$_{CNN}$ + Clf (Ours) | 404$k$ | 89.94 | 92.84  | 92.95  | 93.91  | 76.27  | 93.71  |
|               | Emb$_{300}$ + 5 + Enc$_{CNN}$ + Clf (Ours) | 267$k$ | 90.79 | 93.38  | 93.82  | 94.55  | 77.47  | 94.71  |

Table 4: Precision of selected words for each aspect. Percentage of words corresponds to the ratio of the number of highlighted words to the full review. All models are trained on Filtered Beer.

| Interp.       | Model                                      | Smell | Palate | Appearance |
|---------------|--------------------------------------------|-------|--------|------------|
| Fine-grained  | SVM*                                       | 21.6  | 24.9   | 38.3       | 13%        |
|               | SAA*                                       | 88.4  | 65.3   | 80.6       | 13%        |
|               | SAM (Lei et al., 2016)                     | 95.1  | 80.2   | 96.3       | 14%        |
|               | MASA                                       | 87.0  | 42.8   | 74.5       | 4%         |
|               | MAA                                        | 51.3  | 32.9   | 44.9       | 14%        |
|               | MAM (Ours)                                 | **96.6** | **81.7** | **96.7** | **14%**    |

* The model has been trained separately for each aspect.

Finally, we show that learning soft multi-dimensional masks along training objectives achieves strong predictive results, and using these to create contextualized word embeddings and train a baseline model with, provides the best performance across the three datasets.

4.4 Mask Interpretability

In these experiments, we verify that Multi-Aspect Masker generates induced masks $M_{a1}, ..., M_{aA}$ that, in addition to improving performance, are meaningful and can be interpreted by humans.

4.4.1 Mask Precision

Evaluating justifications that have short and coherent pieces of text is challenging because there is no gold standard provided with reviews. McAuley et al. (2012) provided 994 beer reviews with aspect sentence-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. For both, ours and Lei et al. (2016), we used trained models on beer reviews and extracted a similar number of selected words.

We show that the generated sub-masks $M_{a1}, M_{a2}, M_{a3}$ obtained with Multi-Aspect Masker (MAM) correlate best with human judgment. Table 4 presents the precision of the masks and attentions computed on sentence-level aspect annotations. We reported results of the models in Lei et al. (2016): SVM, the Single Aspect-Attention (SAA) and Single Aspect-Mask (SAM) — trained separately for
MAM model obtains significant higher precisions with an average of each aspect because they find hard focuses only on a couple of words due to the sparseness of the attention (examples in Appendix A.3.1). SVM and attention models perform poorly compared with mask models: especially MASA that focuses only on a couple of words due to the sparseness of the attention (examples in Appendix A.3.1).

Table 5: Average Topic Coherence (NPMI) across different top-$N$ words for each dataset. Each aspect $a_i$ is considered as a topic and the masks (or attentions) are used to compute $P(w|a_i)$.

| Model       | $N = 5$ | $N = 10$ | $N = 15$ | $N = 20$ | $N = 25$ | $N = 30$ | Mean\(^\dagger\) |
|-------------|---------|---------|---------|---------|---------|---------|----------------|
| **Filtered Beer** |         |         |         |         |         |         |                |
| SAM\(^\ddagger\) [Lei et al. 2016] | 0.123   | 0.149   | 0.134   | 0.169   | 0.219   | 0.248   | 0.174         |
| MASA        | 0.024   | 0.059   | 0.159   | 0.200   | 0.271   | 0.325   | 0.173         |
| MAA         | 0.072   | 0.103   | 0.141   | 0.208   | 0.259   | 0.325   | 0.185         |
| MAM (Ours)  | 0.042   | 0.114   | 0.171   | 0.216   | 0.276   | 0.329   | 0.192         |
| **Full Beer** |         |         |         |         |         |         |                |
| SAM\(^\ddagger\) [Lei et al. 2016] | 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         |
| MAM (Ours)  | 0.083   | 0.187   | 0.264   | 0.348   | 0.410   | 0.477   | 0.295         |
| **Hotel**   |         |         |         |         |         |         |                |
| SAM\(^\ddagger\) [Lei et al. 2016] | 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         |
| MAM (Ours)  | 0.134   | 0.251   | 0.349   | 0.496   | 0.641   | 0.724   | 0.432         |

* The model has been trained separately for each aspect.
† Metric that correlates best with human judgment \(\text{(Lau & Baldwin 2016)}\).

Each aspect because they find hard justifications for a single aspect. In comparison to SAM, our MAM model obtains significant higher precisions with an average of +1.13 F1 score. Interestingly, SVM and attention models perform poorly compared with mask models: especially MASA that focuses only on a couple of words due to the sparseness of the attention (examples in Appendix A.3.1).

### 4.5 Mask Coherence

In addition to evaluating masks with human annotations, we computed their semantic interpretability for each dataset. According to \(\text{Aletras & Stevenson (2013); Lau et al. (2014); NPMI (Bouma, 2009)}\), 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 \(\text{Lau & Baldwin (2016)}\) and computed the topic coherence over several cardinalities $N$, and report all the results, as well as the average; the authors claim the mean leads to a more stable and robust evaluation. More details are available in Appendix A.4.

We show that generated masks by MAM obtains the highest mean NPMI and, on average, superior results in all datasets (17 out of 21 cases), while only needing a single training. Results are shown in Table 5. For the Hotel and Full Beer datasets, where reviews reflect the real data distribution, our model significantly outperforms SAM and attention models for $N \geq 20$. For smaller $N$, MAM obtains higher scores in four out of six cases, and for these two, the difference is only below 0.003.

For the controlled-environment Filtered Beer dataset, MAM still performs better for $N \geq 15$, although the differences are smaller, and is beat by SAM for $N \leq 10$. However, SAM obtains poor results in all other cases of all datasets and must be trained as many times as the number of aspects.

We show the top words for each aspect and a human evaluation in Appendix A.4. Generally, our model finds better sets of words among the three datasets compared with other methods.

### 5 Conclusion

In this work, we propose Multi-Aspect Masker, an end-to-end neural network architecture to perform multi-aspect sentiment classification for reviews. Our model predicts aspect sentiments while generating a probabilistic (soft) multi-dimensional mask (one dimension per aspect) simultaneously,
in an unsupervised and multi-task learning manner. We showed that the induced mask is beneficial to guide the model to focus on different parts of the input text and to further improve the sentiment prediction for all aspects. Our evaluation shows that on three datasets, in the beer and hotel domain, our model outperforms strong baselines and generates masks that are: **strong feature predictors, meaningful, and interpretable** compared to attention-based methods and a single-aspect masker.

**ACKNOWLEDGMENTS**

We thank Michaela Benk for proofreading and helpful advice.

**REFERENCES**

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 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, 2015, Conference Track Proceedings [Bahdanau et al. (2015)]. URL http://arxiv.org/abs/1409.0473.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1903–1913, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1216. URL https://www.aclweb.org/anthology/D18-1216.

G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, E. de Castilho, and M. Stede (eds.), Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, pp. 31–40, Tübingen, 2009. Gunter Narr Verlag.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT’s attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4828.

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

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 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, pp. 1606–1615, Denver, Colorado, May–June 2015a. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL https://www.aclweb.org/anthology/N15-1184.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A Smith. 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), pp. 1491–1500, 2015b.

Athanasios Giannakopoulos, Diego Antognini, Claudiu Musat, Andreea Hossmann, and Michael Baesischyl. Dataset construction via attention for aspect term extraction with distant supervision. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 373–380. IEEE, 2017.

Aurélie Herbelot and Eva Maria Vecchi. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 22–32, 2015.
Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* Jang et al. (2017). URL https://openreview.net/forum?id=rKE3y85ee

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. Visualizing and understanding recurrent networks. *CoRR*, abs/1506.02078, 2015. URL http://arxiv.org/abs/1506.02078

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* Kingma & Ba (2015). URL http://arxiv.org/abs/1412.6980

Jey Han Lau and Timothy Baldwin. 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: Human Language Technologies*, pp. 483–487, 2016.

Jey Han Lau, David Newman, and Timothy Baldwin. 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*, pp. 550–539, 2014.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://www.aclweb.org/anthology/D16-1011

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 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*, pp. 681–691, 2016a.

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

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

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* Lin et al. (2017). URL https://openreview.net/forum?id=BJC_ jUqxe

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* Maddison et al. (2017). URL https://openreview.net/forum?id=S1jE5L5g1
Iain J Marshall, Joël Kuiper, and Byron C Wallace. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association, 23(1):193–201, 2015.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In International Conference on Machine Learning, pp. 1614–1623, 2016.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, ICDM ’12, pp. 1020–1025, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-0-7695-4905-7. doi: 10.1109/ICDM.2012.110.

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

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

Claudiu Cristian Musat, Yizhong Liang, and Boi Faltings. Recommendation using textual opinions. In Twenty-Third International Joint Conference on Artificial Intelligence, 2013.

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

Nikolaos Pappas and Andrei Popescu-Belis. 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 (EMNLP), pp. 455–466, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

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

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

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

Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 783–792. ACM, 2010.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, pp. 1480–1489, 2016.

Yichun Yin, Yangqiu Song, and Ming Zhang. Document-level multi-aspect sentiment classification as machine comprehension. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2044–2054, 2017.
Ye Zhang, Iain Marshall, and Byron C. Wallace. Rationale-augmented convolutional neural networks for text classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 795–804, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1076.
A Appendix

A.1 Descriptive Statistics of the Datasets

Table 6: Statistics of the multi-aspect review datasets. Full Beer and Hotel represent real-world beer and hotel reviews respectively. Filtered Beer contains a subset of beer reviews with assumptions, leading to a more straightforward and unrealistic dataset.

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

A.2 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 & 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 & Hinton], 2010). We used the 200-dimensional pre-trained word embeddings of [Lei et al.], (2016) for beer reviews. For the hotel domain, we trained word2vec ([Mikolov et al.], 2013) on a large collection of hotel reviews with an embedding size of 300.

We used dropout ([Srivastava et al.], 2014) of 0.1, clipped the gradient norm at 1.0 if higher, added L2-norm regularizer with a regularization factor of 10^-6 and trained using early stopping with a patience of three iterations. We used Adam ([Kingma & Ba], 2015) for training with a learning rate of 0.001, β1 = 0.9, and β2 = 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 λsel = 0.03, λcont = 0.04, and λp = 0.11 for the Filtered Beer dataset; λsel = 0.03, λcont = 0.03, and λp = 0.15 for the Full Beer dataset; and λsel = 0.02, λcont = 0.02 and λp = 0.10 for the Hotel dataset. We ran all experiments for a maximum of 50 epochs with a batch-size of 256 and a Titan X GPU. For the model of [Lei et al.], (2016), we reused the code from the authors.

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-Aspect Masker model (MAM), the Single Aspect-Mask method (SAM) of [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 4.2). 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 Filtered Beer, Full Beer and Hotel dataset are shown below.
### A.3.1 Filtered Beer Dataset

| Multi Aspect-Masks (Ours) | Single Aspect-Mask (Lei et al., 2016) |
|---------------------------|---------------------------------------|
| **Appearance**<br> ruby red brown in color. fluffy off white single<br> finger head settles down to a thin cap coating thin lacing all over the sides on the glass. 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. my wife is a big brown fan, so i'll definitely be grabbing this one for her again. a solid beer for any time of the year. served: in a standard pint glass. | **Appearance**<br> ruby red brown in color. fluffy off white single<br> finger head settles down to a thin cap coating thin lacing all over the sides on the glass. 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. my wife is a big brown fan, so i'll definitely be grabbing this one for her again. a solid beer for any time of the year. served: in a standard pint glass. |

| Multi Aspect-Attentions | Multi Aspect-Sparse-Attentions |
|-------------------------|-------------------------------|
| **Appearance**<br> ruby red brown in color. fluffy off white single<br> finger head settles down to a thin cap coating thin lacing all over the sides on the glass. 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. my wife is a big brown fan, so i'll definitely be grabbing this one for her again. a solid beer for any time of the year. served: in a standard pint glass. | **Appearance**<br> ruby red brown in color. fluffy off white single<br> finger head settles down to a thin cap coating thin lacing all over the sides on the glass. 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. my wife is a big brown fan, so i'll definitely be grabbing this one for her again. a solid beer for any time of the year. served: in a standard pint glass. |

Figure 3: Our model MAM highlights all the words corresponding to 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 the Filtered Beer dataset does not include it in the aspect set.
Under review as a conference paper at ICLR 2020

Figure 4: MAM 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.
A.3.2 Full Beer Dataset

| Appearance | Smell | Palate | Taste |
|------------|-------|--------|-------|
| 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 " 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 pumpkin, 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-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 " 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 pumpkin, 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 5: MAM can identify accurately what parts of the review describe each aspect. Due to the high imbalance and correlation, MAA provides very noisy labels, while MASA highlights only a few important words. We can see that SAM is confused and performs a poor selection.
Appearance Smell Palate Taste

Multi Aspect-Masks (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-Mask (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-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 6: On a short review, MAM achieves near-perfect annotations, while SAM highlights only two words where one is ambiguous with respect to four aspects. MAA mixes between the aspect Appearance and Smell. MASA identifies some words but lacks coverage.
A.3.3 Hotel Dataset

| Service | Cleanliness | Value | Location | Room |
|---------|-------------|-------|----------|------|
| Multi Aspect-Masks (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 n’t 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-Mask (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 n’t 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 n’t 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 n’t 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 7: MAM emphasizes consecutive words, identifies important 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 important key-phrases, and labels are noisy when the confidence is not high. MASA provides noisier tags than MAA.
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 scratches 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 ! fryps - 2euro would go back again , but mayby stay somewhere else closer to the beach ( sol pelicanos etc ) ... this hotel is next to an alley called ' muggers alley '.

Multi-Aspect Masker (Ours)
A.4 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 $w_{a_i}^N$, the Normalized Pointwise Mutual Information (Bouma, 2009) coherence score is:

$$NPMI(w_{a_i}^N) = \sum_{j=2}^{N} \sum_{k=1}^{j-1} \frac{\log \frac{P(w_k^{a_i}, w_j^{a_i})}{P(w_k^{a_i})P(w_j^{a_i})}}{\log \frac{P(w_k^{a_i})}{P(w_k^{a_i})} - \log \frac{P(w_j^{a_i})}{P(w_j^{a_i})}}$$

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 averages across all aspect words for each word $w$: $b_w = \frac{1}{|A|} \sum_{i=1}^{|A|} P(w|a_i)$, which generates 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 subsequent tables 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.

Table 7: Top ten words for each aspect from the Filtered Beer dataset, learned by various models. Red denotes intruders according to two human annotators. For the three aspects, MAM has only one word considered as an intruder, followed by MASA with SAM (two) and MAA (six).

| 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 8: Top ten words for each aspect from the Full 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                                      |
|-----------|---------------------------------------------------|
| Appearance |                                                  |
| SAM       | nothing beautiful lager nice average              |
|           | 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 thin                                        |
| Smell     |                                                  |
| SAM       | faint nice mild light slight complex              |
|           | good wonderful grainy great                       |
| MASA      | aroma hops nose chocolate malt citrus fruit      |
|           | smell fruits                                      |
| MAA       | taste hints hint 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                |
| MAM (Ours)| carbonation medium mouthfeel body smooth bodied  |
|           | drinkability creamy light                         |
| 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                  |
| Room      |                                                  |

Table 9: 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, where MAM does not have an intruder.

| 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 smoking 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                       |
| 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               |
|           | 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                                       |
A.5 Baseline Architectures

![Diagram of Baseline model Emb + Enc CNN + Clf.](image1)

![Diagram of Baseline model Emb + Enc CNN + A Shared + Clf.](image2)

![Diagram of Baseline model Emb + Enc LSTM + A Shared + Clf.](image3)

![Diagram of Baseline model Emb + Enc LSTM + A Aspect-wise + Clf.](image4)

Figure 9: Baseline model Emb + Enc CNN + Clf.

Figure 10: Baseline model Emb + Enc CNN + A Shared + Clf.

Figure 11: Baseline model Emb + Enc LSTM + A Shared + Clf.

Figure 12: Baseline model Emb + Enc LSTM + A [Sparse] Aspect-wise + Clf. Attention is either additive or sparse.