Multi-Dimensional Explanation of Target Variables from Documents

Diego Antognini,1 Claudiu Musat,2 Boi Faltings 1
1Ecole Polytechnique Fédérale de Lausanne, Switzerland
2Swisscom, Switzerland
{diego.antognini,claudiu.musat,boi.faltings}@{epfl.ch, swisscom.com}

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

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

Perhaps the simplest means of explaining predictions of complex models is by selecting relevant input features. Prior work includes various methods to find relevant words in the text input to predict the target variable of a document. Attention mechanisms (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015) model the word selection by a conditional importance distribution over the inputs, used as explanations to produce a weighted context vector for downstream modules. However, their reliability has been questioned (Jain and Wallace 2019; Pruthi et al. 2020). Another line of research includes rationale generation methods (Lundberg and Lee 2017; Li, Monroe, and Jurafsky 2016; Lei, Barzilay, and Jaakkola 2016). If the selected text input features are short and concise – called a rationale or mask – and suffice on their own to yield the prediction, it can potentially be understood and verified against domain knowledge (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2019). Specifically, these rationale generation methods have been recently proposed to provide such explanations alongside the prediction. Ideally, a good rationale should yield the same or higher performance as using the full input.

The key motivation of our work arises from the limitations of the existing methods. First, the attention mechanisms induce an important distribution over the inputs, but the resulting explanation consists of many short and noisy word sequences (Figure 1). In addition, the rationale generation methods produce coherent explanations, but the rationales are based on a binary selection of words, leading to the following shortcomings: 1. they explain only one target variable, 2. they make a priori assumptions about the data, and 3. they make it difficult to capture ambiguities in the text. Regarding the first...
shortcoming, rationales can be multi-faceted by definition and involve support for different outcomes. If that is the case, one has to train, tune, and maintain one model per target variable, which is impractical. For the second, current models are prone to pick up spurious correlations between the input features and the output. Therefore, one has to ensure that the data have low correlations among the target variables, although this may not reflect the real distribution of the data. Finally, regarding the last shortcoming, a strict assignment of words as rationales might lead to ambiguities that are difficult to capture. For example, in an hotel review that states “The room was large, clean, and close to the beach.”, the word “room” refers to the aspects Room, Cleanliness, and Location. All these limitations are implicitly related due to the non-probabilistic nature of the mask. For further illustrations, see Figure 3 and the appendices.

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

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

**Related Work**

**Interpretability**

Developing interpretable models is of considerable interest to the broader research community; this is even more pronounced with neural models (Kim, Shah, and Doshi-Velez 2015; Doshi-Velez and Kim 2017). There has been much work with a multitude of approaches in the areas of analyzing and visualizing state activation (Karpathy, Johnson, and Li 2015; Li et al. 2016; Montavon, Samek, and Müller 2018), attention weights (Jain and Wallace 2019; Serrano and Smith 2019; Pruthi et al. 2020), and learned sparse and interpretable word vectors (Faruqui et al. 2015b,a; Herbelot and Vecchi 2015). Other works interpret black box models by locally fitting interpretable models (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). (Li, Monroe, and Jurafsky 2016) proposed erasing various parts of the input text using reinforcement learning to interpret the decisions. However, this line of research aims at providing post-hoc explanations of an already-trained model. Our work differs from these approaches in terms of what is meant by an explanation and its computation. We defined an explanation as one or multiple text snippets that – as a substitute of the input text – are sufficient for the predictions.

**Attention-based Models**

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

**Rationale Models**

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

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

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

The Multi-Target Masker (MTM)

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

Model Overview

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

$$P(M | X) = \prod_{t=1}^{L} P(M^t | x^t) = \prod_{t=1}^{L} \prod_{i=0}^{T} P(m^t_i | x^t)$$

Because we have categorical distributions, we cannot directly sample $P(M^t | x^t)$ and backpropagate the gradient through this discrete generation process. Instead, we model the variable $m^t_i$ using the straight-through gumbel-softmax (Jang, Gu, and Poole 2017; Maddison, Mnih, and Teh 2017) to approximate sampling from a categorical distribution. We model the parameters of each Gumbel-Softmax distribution $M^t$ with a single-layer feed-forward neural network followed by applying a log softmax, which induces the log-probabilities of the $i^{th}$ distribution: $\omega_i = \log(\text{softmax}(W h^t + b))$, $W$ 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 $\tau$, which dictates the peakiness of the relevance distributions. In our case, we keep the temperature low to enforce the assumption that each word is relevant about one or two targets. Note that compared to attention mechanisms, the word importance is a probability distribution over the targets $\sum_{i=0}^{T} P(m^t_i | x^t) = 1$ instead of a normalization over the sequence length $\sum_{t=1}^{L} P(t^t | x^t) = 1$.

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

$$M_{t_i} = P(m^t_1 | x^t), P(m^t_2 | x^t), \ldots, P(m^t_L | x^t)$$

To integrate the word importance of the induced sub-masks $M_{t_i}$ within the model, we weight the word embeddings by their importance towards a target variable $t_i$ such that $E_{t_i} = E \cap M_{t_i} = e_1 \cdot P(m^t_1 | x^t), e_2 \cdot P(m^t_2 | x^t), \ldots, e_L \cdot P(m^t_L | x^t)$. Thereafter, each modified embedding $E_{t_i}$ is fed into the encoder block. Note that $E_{t_0}$ is ignored because $M_{t_0}$ only serves to absorb probabilities of words that are insignificant.

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

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

**Enabling the Interpretability of Masks**

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

$$\ell_{\text{pred}} = \sum_{i=1}^{T} \ell_{\text{cross-entropy}}(y_{t_i}, \hat{y}_{t_i})$$

---

1Our method is easily adapted for regression problems.

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

3If $P(m^t_i | x^t) \approx 1.0$, it implies $\sum_{i=1}^{T} P(m^t_i | x^t) \approx 0$ and consequently, $e_i^t \approx 0$ for $i = 0, \ldots, T$. 

12509
However, training MTM to optimize \( \ell_{\text{pred}} \) will lead to meaningless sub-masks \( M_i \) because the model tends to focus on certain words. Consequently, we guide the model to produce long, meaningful word sequences, as shown in Figure 1. We propose two regularizers to control the number of selected words and encourage consecutive words to be relevant to the same target. For the first term, we calculate the probability \( p_{\text{sel}} \) of tagging a word as relevant to any target as follows:

\[
p_{\text{sel}} = \frac{1}{L} \sum_{\ell=1}^{L} (1 - P(m_{\ell|0}^\ell|x^\ell))
\]  

(4)

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

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

(5)

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

\[
p_{\text{dis}} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{|M_\ell - M_{\ell-1}|}{A + 1}
\]

(6)

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

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

**Experiments**

We assess our model in two dimensions: the quality of the explanations, obtained from the masks, and the predictive performance. Following previous work (Lei, Barzilay, and Jaakkola 2016; Chang et al. 2020), we use sentiment analysis as a demonstration use case, but we extend it to the multi-aspect case. However, we are interested in learning rationales for every aspect at the same time without any prior assumption on the data, where aspect ratings can be highly correlated.

We first measure the quality of the induced rationales using human aspect sentence-level annotations and an automatic topic model evaluation method. In the second set of experiments, we evaluate MTM on the multi-aspect sentiment classification task in two different domains.

**Experimental Details**

The review encoder was either a bi-directional recurrent neural network using LSTM (Hochreiter and Schmidhuber 1997) with 50 hidden units or a multi-channel text convolutional neural network, similar to (Kim, Shah, and Doshi-Velez 2015), with 3-, 5-, and 7-width filters and 50 feature maps per filter. Each aspect classifier is a two-layer feedforward neural network with a rectified linear unit activation function (Nair and Hinton 2010). We used the 200-dimensional pre-trained word embeddings of (Lei, Barzilay, and Jaakkola 2016) for beer reviews. For the hotel domain, we trained word2vec (Mikolov et al. 2013) on a large collection of hotel reviews (Antognini and Faltings 2020) with an embedding size of 300. We used a dropout (Srivastava et al. 2014) of 0.1, clipped the gradient norm at 1.0, added a L2-norm regularizer with a factor of \( 10^{-6} \), and trained using early stopping. We used Adam (Kingma and Ba 2015) with a learning rate of 0.001. The temperature \( \tau \) for the Gumbel-Softmax distribution was fixed at 0.8. The two regularizers and the prior of our model were \( \lambda_{\text{sel}} = 0.03 \), \( \lambda_{\text{cont}} = 0.03 \), and \( \lambda_p = 0.15 \) for the Beer dataset and \( \lambda_{\text{sel}} = 0.02 \), \( \lambda_{\text{cont}} = 0.02 \), and \( \lambda_p = 0.10 \) for the Hotel one. We ran all experiments for a maximum of 50 epochs with a batch-size of 256. We tuned all models on the dev set with 10 random search trials.

**Datasets**

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

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

Table 1: Statistics of the multi-aspect review datasets. Both datasets have high correlations between aspects.
We first verify whether the inferred rationales of $\text{MTM}$ are meaningful and interpretable, compared to the other models.

**Precision.** Evaluating explanations that consist of coherent pieces of text is challenging because there is no gold standard for reviews. (McAuley, Leskovec, and Jurafsky 2012) have provided 994 beer reviews with sentence-level aspect annotations (although our model computes masks at a finer level). Each sentence was annotated with one aspect label, indicating what aspect that sentence covered. We evaluate the precision of the words selected by each model, as in (Lei, Barzilay, and Jaakkola 2016). We use trained models on the $\text{Beer}$ dataset and extracted a similar number of selected words for a fair comparison. We also report the results of the models from (Lei, Barzilay, and Jaakkola 2016); $\text{NB-SVM}$, the Single-Aspect Attention and Masker ($\text{SAA}$ and $\text{SAM}$, respectively); they use the separate decorrelated train sets for each aspect because they compute hard masks.\(^3\)

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

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

The results are shown in Table 3. We show that the computed masks by $\text{MTM}$ lead to the highest mean NPMI and, on average, 20% superior results in both datasets, while only needing a single training. Our $\text{MTM}$ model significantly outperforms $\text{SAM}$ and the attention models ($\text{MASA}$ and $\text{MAA}$) for $N \geq 20$ and $N = 5$. For $N = 10$ and $N = 15$, $\text{MTM}$
obtains higher scores in two out of four cases (+.033 and +.009). For the other two, the difference was below .003. SAM obtains poor results in all cases.

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

Multi-Aspect Sentiment Classification

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

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

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

NB-SVM, which offers fine-grained interpretability and was trained separately for each aspect, significantly underperforms when compared to BASE and, surprisingly, to SENT. As shown in Table 1, the sentiment correlation between any pair of aspects of the Beer dataset is on average 71.8%. Therefore, by predicting the sentiment of one aspect correctly, it is likely that other aspects share the same polarity. We suspect that the linear model NB-SVM cannot capture the correlated relationships between aspects, unlike the non-linear (neural) models that have a higher capacity. The shared attention models perform better than BASE but provide only coarse-grained interpretability. SAM is outperformed by all the models except SENT, BASE, and NB-SVM.

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

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

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

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

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

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

**Conclusion**

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

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

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

<table>
<thead>
<tr>
<th>Interp.</th>
<th>Model</th>
<th>Params</th>
<th>F1 Scores</th>
<th>F1 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Macro</td>
<td>A1</td>
</tr>
<tr>
<td>None</td>
<td>SENT</td>
<td>309k</td>
<td>85.91</td>
<td>89.98</td>
</tr>
<tr>
<td>BASE</td>
<td>Emb300 + EncCNN + Clf</td>
<td>263k</td>
<td>90.30</td>
<td>92.91</td>
</tr>
<tr>
<td>Coarse-grained SAA</td>
<td>Emb300 + EncCNN + Ashared + Clf</td>
<td>310k</td>
<td>90.12</td>
<td>92.73</td>
</tr>
<tr>
<td></td>
<td>Emb300 + EncCNN + AAspect-wise + Clf</td>
<td>270k</td>
<td>88.22</td>
<td>91.13</td>
</tr>
<tr>
<td>NM-SVM</td>
<td>(Wang and Manning 2012)</td>
<td>5 - 309k</td>
<td>87.17</td>
<td>90.04</td>
</tr>
<tr>
<td>SAM</td>
<td>(Lei, Barzilay, and Jaakkola 2016)</td>
<td>5 - 824k</td>
<td>87.52</td>
<td>91.48</td>
</tr>
<tr>
<td>MASA</td>
<td>Emb300 + EncLSTM + AAspect-wise + Clf</td>
<td>1010k</td>
<td>90.23</td>
<td>93.11</td>
</tr>
<tr>
<td>MAA</td>
<td>Emb300 + EncLSTM + AAspect-wise + Clf</td>
<td>1010k</td>
<td>90.21</td>
<td>92.84</td>
</tr>
<tr>
<td>MTM</td>
<td>Emb300 + Masker + EncCNN + Clf (Ours)</td>
<td>404k</td>
<td>89.94</td>
<td>92.84</td>
</tr>
<tr>
<td>MTM C</td>
<td>Emb300 + Masker + EncCNN + Clf (Ours)</td>
<td>267k</td>
<td>90.79</td>
<td>93.38</td>
</tr>
</tbody>
</table>

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


