A Sparse attention

A natural way to get a sparse attention distribution is by using the **sparsemax transformation** (Martins and Astudillo, 2016), which computes an Euclidean projection of the score vector onto the probability simplex $\triangle^n := \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} \geq \mathbf{0},\ \mathbf{1}^\top \mathbf{p} = 1\}$, or, more generally, the α**-entmax transformation** (Peters et al., 2019):

$$\alpha\text{-entmax}(\mathbf{s}) := \arg\max_{\mathbf{p} \in \triangle^n} \mathbf{p}^\top \mathbf{s} + H_\alpha(\mathbf{p}), \tag{2}$$

where H_α is a generalization of the Shannon and Gini entropies proposed by Tsallis (1988), parametrized by a scalar $\alpha \geq 1$:

$$H_\alpha(\mathbf{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \sum_j (p_j - p_j^\alpha), & \alpha \neq 1 \\ -\sum_j p_j \log p_j, & \alpha = 1. \end{cases} \tag{3}$$

Setting $\alpha = 1$ recovers the softmax function, while for any value of $\alpha > 1$ this transformation can return a sparse probability vector. Letting $\alpha = 2$, we recover sparsemax. A popular choice is $\alpha = 1.5$, which has been successfully used in machine translation and morphological inflection applications (Peters et al., 2019).

B Data statistics and preparation

We used four datasets for text classification: SST,[7] IMDB,[8] AgNews,[9] and Yelp.[10] One dataset for NLI: SNLI,[11] along with its extended version (eSNLI[12]) which includes human-annotated explanations of the entailment relations (Camburu et al., 2018). And the EN→DE IWSLT 2017 dataset for machine translation (Cettolo et al., 2017).[13] Table 5 shows statistics for each dataset.

NAME	# TRAIN	# TEST	AVG. TOKENS	# CLASSES
SST	6920	1821	19	2
IMDB	25K	25K	280	2
AgNews	115K	20K	38	2
Yelp	5.6M	1M	130	5
SNLI	549K	9824	14 / 8	3
IWSLT	206K	2271	20 / 18	134,086

Table 5: Dataset statistics. The average number of tokens for SNLI is related to the premise and hypothesis, and for IWSLT to the source and target sentences.

For AgNews, we considered the binary case of World vs Business articles. Although the selected datasets are the same as in previous works (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), the training and test set might differ. For SST, IMDB, IWSLT and SNLI we used the standard splits, but for AgNews and Yelp we randomly split the dataset, leaving 85% for training and 15% for test. Moreover, for IMDB, AgNews and Yelp we randomly selected 10%, 15% and 15% of examples from the training set to be used as validation data, respectively.

C Computing infrastructure

Our infrastructure consists of 4 machines with the specifications shown in Table 6. The machines were used interchangeably, and all experiments were executed in a single GPU. Despite having machines with different specifications, we did not observe large differences in the execution time of our models across different machines.

[7] https://nlp.stanford.edu/sentiment/
[8] https://ai.stanford.edu/~amaas/data/sentiment/
[9] https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
[10] https://www.yelp.com/dataset/
[11] https://nlp.stanford.edu/projects/snli/
[12] https://github.com/OanaMariaCamburu/e-SNLI
[13] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

#	GPU	CPU
1.	4 × Titan Xp - 12GB	16 × AMD Ryzen 1950X @ 3.40GHz - 128GB
2.	4 × GTX 1080 Ti - 12GB	8 × Intel i7-9800X @ 3.80GHz - 128GB
3.	3 × RTX 2080 Ti - 12GB	12 × AMD Ryzen 2920X @ 3.50GHz - 128GB
4.	3 × RTX 2080 Ti - 12GB	12 × AMD Ryzen 2920X @ 3.50GHz - 128GB

Table 6: Computing infrastructure.

D Classifiers experimental setup (Table 2)

We chose our classifiers so that they are close to the models used by related works (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings et al., 2019). For all models, we calculated their accuracy on the dev set after each epoch. At the end of training we selected the model with the best validation accuracy. We experimented with two classes of classifiers: a simple RNN with attention as in Jain and Wallace (2019); Wiegreffe and Pinter (2019); and the rationalizer models of Lei et al. (2016) and Bastings et al. (2019) which sample binary masks from Bernoulli and HardKuma distributions, respectively.

D.1 RNNs with attention

For the text classification experiments, each input word x_i is mapped to 300D-pretrained GloVe embeddings (Pennington et al., 2014) from the 840B release,[14] kept frozen, followed by a bidirectional LSTM layer (BiLSTM) resulting in vectors $\mathbf{h}_1, \ldots, \mathbf{h}_n$. We score each of these vectors using the additive formulation of Bahdanau et al. (2015), applying an attention transformation to convert the resulting scores $\mathbf{s} \in \mathbb{R}^n$ to a probability distribution $\pi \in \triangle^n$. We use this to compute a contextual vector $\mathbf{c} = \sum_{i=1}^n \pi_i \mathbf{h_i}$, which is fed into the output softmax layer that predicts $\hat{y}$. For NLI, the input x is a pair of sentences (a premise and an hypothesis), and the classifier C is similar to the the above, but with two independent BiLSTM layers, one for each sentence. In the attention layer, we use the last hidden state of the hypothesis as the query and the premise vectors as keys.

We used the AdamW (Loshchilov and Hutter, 2019) optimizer for all experiments. We tuned two hyperparameters: learning rate within $\{0.003, \mathbf{0.001}, 0.0001\}$, and l_2 regularization within $\{0.01, 0.001, \mathbf{0.0001}, 0\}$. We picked the best configuration by doing a grid search and by taking into consideration the accuracy on the validation set (selected values in bold). Table 7 shows all hyperparameters set for training.

HYPERPARAM.	SST	IMDB	AGNEWS	YELP	SNLI
Word embeddings size	300	300	300	300	300
BiLSTM hidden size	128	128	128	128	128
Merge BiLSTM states	concat	concat	concat	concat	concat
Batch size	8	16	16	128	32
Number of epochs	10	10	5	5	10
Early stopping patience	5	5	3	3	5
Learning rate	0.001	0.001	0.001	0.001	0.001
ℓ_2 regularization	0.0001	0.0001	0.0001	0.0001	0.0001

Table 7: RNNs training hyperparameters for text classification and NLI datasets.

D.2 Bernoulli and HardKuma

We used the implementation of Bastings et al. (2019),[15] which includes a reimplementation of the generator-encoder model from (Lei et al., 2016). The model used for text classification is a RNN-based generator followed by a RNN-based encoder, whereas for NLI is a decomposable attention classifier from (Parikh et al., 2016), for which only the HardKuma implementation was available. In order to faithfully compare the frameworks, we adapted the HardKuma code and implemented a Bernoulli version of the

[14] http://nlp.stanford.edu/data/glove.840B.300d.zip
[15] https://github.com/bastings/interpretable_predictions

same classifier, taking into consideration the sparsity and fused-lasso loss penalties, and the deterministic strategy used during test time. For simplicity, we used the independent variant of the generator of Lei et al. (2016). Table 8 lists only the hyperparameters that we set during training. We refer to the original work of Bastings et al. (2019) to see all other hyperparamers, for which we kept the default values.

HYPERPARAM.	SST	IMDB	AGNEWS	YELP	SNLI
Latent selection (HardKuma)	0.3	0.1	0.3	0.3	0.1
Sparsity penalty (Bernoulli)	0.01	0.001	0.01	0.01	0.0003
Lasso penalty	0	0	0	0	0
Batch size	25	25	25	256	64
Number of epochs	25	25	25	10	100
Early stopping patience	5	5	5	5	100
Learning rate	0.0002	0.0002	0.0002	0.001	0.0002
ℓ_2 regularization	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-6}

Table 8: Rationalizer models training hyperparameters for text classification and NLI datasets.

D.3 Validation set results and model statistics

Table 9 shows the accuracy of each classifier on the validation set, their number of trainable parameters and the average training time per epoch.

	SST			IMDB			AGNEWS		
CLF.	# P	t	ACC	# P	t	ACC	# P	t	ACC
C	474K	10s	85.32	474K	2m	95.64	474K	2m	98.09
C_{ent}	474K	10s	84.29	474K	2m	95.84	474K	2m	98.54
C_{sp}	474K	10s	84.17	474K	2m	95.44	474K	2m	98.51
C_{bern}	1.1M	15s	80.16	1.1M	2m	87.40	1.1M	2m	96.26
C_{hk}	1.1M	15s	84.40	1.1M	2m	91.84	1.1M	2m	96.74

Table 9: Classifier results on the validation set and model statistics. # P is the number of trainable parameters, and is t the average training time per epoch.

	YELP			SNLI		
CLF.	# P	t	ACC	# P	t	ACC
C	474K	3h	77.03	998K	4m	78.74
C_{ent}	474K	3h	76.72	998K	4m	79.38
C_{sp}	474K	3h	76.84	998K	4m	79.69
C_{bern}	1.1M	5h	69.99	382K	2m	79.79
C_{hk}	1.1M	5h	74.29	462K	2m	86.04

Table 10: Continuation of Table 9.

E Communication experimental setup (Table 3)

Training the communication under our framework consists on training a layperson L on top of explanations (message) produced by E about C's decision. With the exception of the explainer E trained jointly with L, none of the other explainers have trainable parameters. Therefore, in these cases, the communication between E and L consists only on training L. For all models, we calculated its CSR on the dev set after each epoch. At the end of training we selected the model with the best validation CSR. Table 11 shows the communication hyperparameters. Note that for SNLI we still need to train a BiLSTM to encode the hypothesis.

HYPERPARAM.	SST	IMDB	AGNEWS	YELP	SNLI
Word embeddings size	-	-	-	-	300
BiLSTM hidden size	-	-	-	-	128
Merge BiLSTM states	-	-	-	-	concat
Batch size	16	16	16	112	64
Number of epochs	10	10	10	5	10
Early stopping patience	3	3	3	3	3
Learning rate	0.001	0.001	0.001	0.003	0.001
ℓ_2 regularization	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}

Table 11: Communication hyperparameters for text classification and NLI datasets.

E.1 Validation set results and model statistics

Table 12 shows the CSR and ACC_L for each explainer on the validation set, the number of trainable parameters of L and the average training time per epoch.

	SST				IMDB				AGNEWS			
EXPLAINER	# P	t	CSR	ACC_L	# P	t	CSR	ACC_L	# P	t	CSR	ACC_L
Random	38K	10s	63.76	62.84	247K	1m	61.36	61.24	120K	2m	85.26	84.58
Erasure	38K	10s	81.88	79.82	247K	2m	94.00	91.40	120K	3m	98.41	96.98
Top-k gradient	38K	10s	76.72	75.57	247K	1m	91.88	89.52	120K	2m	98.23	96.97
Top-k softmax	38K	20s	84.29	80.62	247K	1m	96.60	93.60	120K	2m	98.54	97.14
Top-k 1.5-entmax	38K	20s	85.44	80.28	247K	1m	97.88	94.92	120K	2m	98.22	97.37
Top-k sparsemax	38K	20s	85.44	81.54	247K	1m	96.76	93.32	120K	2m	96.46	95.72
Select. 1.5-entmax	38K	10s	85.55	80.62	247K	1m	97.44	94.56	120K	1m	98.30	97.41
Select. sparsemax	38K	10s	85.44	81.54	247K	1m	97.04	93.36	120K	1m	96.46	95.72
Bernoulli	38K	5s	84.75	78.21	247K	1m	91.80	87.36	120K	1m	97.12	94.82
HardKuma	38K	5s	87.50	81.76	247K	1m	95.36	91.20	120K	1m	97.38	96.05

Table 12: Communication results on the validation set and explainer statistics. # P is the number of trainable parameters, and is t the average training time per epoch.

	YELP				SNLI			
EXPLAINER	# P	t	CSR	ACC_L	# P	t	CSR	ACC_L
Random	1.8M	3h	52.55	48.21	560K	9m	31.04	33.11
Erasure	1.8M	4h	79.63	69.59	560K	10m	78.72	70.60
Top-k gradient	1.8M	3h	71.81	63.59	560K	10m	77.55	69.41
Top-k softmax	1.8M	3h	81.49	70.67	560K	9m	79.10	70.95
Top-k 1.5-entmax	1.8M	3h	82.80	71.31	560K	9m	80.30	73.57
Top-k sparsemax	1.8M	3h	82.97	71.46	560K	9m	83.25	75.34
Select. 1.5-entmax	1.8M	2h	82.90	70.99	560K	6m	77.46	71.66
Select. sparsemax	1.8M	2h	84.67	72.25	560K	6m	82.33	75.11
Bernoulli	1.8M	2h	84.93	66.77	560K	2m	75.75	68.61
HardKuma	1.8M	2h	87.43	71.57	560K	3m	75.10	71.10

Table 13: Continuation of Table 12.

F Joint E and L setup

F.1 Communication

According to §4.2, in this model we have two set of parameters to train, one for the explainer E and other for the layperson L, whereas the classifier is a frozen model that we want to explain. Here, we set C as the RNN with softmax classifier (see §2). We design E with the same architecture of the RNNs with attention from §D.1 but without a final output layer, and L have the same architecture as the laypersons in §5. In short, the architecture of E is composed of: (i) embedding layer; (ii) BiLSTM; (iii) attention mechanism. As before, the message is constructed with the words extracted from the attention mechanism.

We use sparsemax attention during training to ensure end-to-end differentiability, and we recover the top-k attended words during test time. We used $k = 5$ for IMDB and $k = 4$ for SNLI in all experiments. In order to encourage faithful explanations, we set $h = \frac{1}{L}\sum_i C_{\text{RNN}}(x_i)$ and $\tilde{h} = \frac{1}{L}\sum_i \text{FFN}(E_{\text{RNN}}(x_i))$, where FFN is a simple feed-forward layer, and $C_{\text{RNN}}(x_i)$ and $E_{\text{RNN}}(x_i)$ are the BiLSTM states from the classifier and the explainer, respectively. In other words, we are approximating the average of the BiLSTM states of C and E. We set $\lambda = 1$ and $\beta = 0.2$ and used the same hyperparameters as in Table 11. The list of stopwords used in our experiments contains 127 English words extracted from NLTK.

F.2 Analysis of β

A potential problem of this model is for the two agents to agree on a trivial protocol, ensuring a high CSR even with bad quality explanations (e.g. punctuations or stopwords). Besides preventing stopwords to be in the message,[16] we set a different probability β of the explainer accessing the predictions of the classifier $\hat{y}$. Intuitively, these strategies should encourage explanations to have higher quality. One way to quantitatively access the quality of the explanations is by aggregating the relative frequencies of each selected word in the validation set, and calculating its Shannon's entropy. If the entropy is low, then the explanations have a high number of repetitions and the explainers are focusing on a very small subset of words, denoting a trivial protocol. To check for a reasonable entropy score that resembles a good quality explanation, we investigate the entropy of the other explainers, for which we had confirmed their quality via human evaluation.

In order to see the impact of β, we carried an experiment with increasing values of β and looked at the CSR, ACC_L and the entropy (H) of the generated explanations. Results are shown in Table 14 for each explainer on IMDB and SNLI.

		IMDB			SNLI		
CLF.	EXPLAINER	H	CSR	ACC_L	H	CSR	ACC_L
C	Random	9.13	59.20	58.92	8.21	31.04	33.11
C	Erasure	9.40	96.32	93.48	9.75	78.72	70.60
C	Top-k gradient	9.49	85.84	83.72	9.39	77.55	69.41
C	Top-k softmax	9.38	94.44	91.84	9.76	78.66	71.00
C_{ent}	Top-k 1.5-entmax	9.62	95.20	93.36	9.54	80.30	73.57
C_{sp}	Top-k sparsemax	9.56	95.28	92.56	8.79	83.25	75.34
C_{ent}	Select. 1.5-entmax	10.76	97.44	94.56	8.49	77.46	71.66
C_{sp}	Selec. sparsemax	10.41	97.04	93.36	8.38	82.33	75.11
C_{bern}	Bernoulli	10.66	91.88	87.36	8.27	75.75	68.61
C_{hk}	HardKuma	11.38	95.36	91.20	9.93	75.10	71.10
-	Human highlights	-	-	-	8.72	87.97	87.97
C	Joint E and L ($\beta = 0.0$)	6.16	93.04	90.84	9.81	80.74	72.38
C	Joint E and L ($\beta = 0.2$)	6.05	98.52	94.56	9.81	93.44	77.20
C	Joint E and L ($\beta = 0.5$)	5.63	99.68	95.64	9.45	95.81	77.54
C	Joint E and L ($\beta = 1.0$)	3.72	99.92	95.56	9.01	97.49	77.23

Table 14: Entropy of the explanations for all explainers on the validation set of IMDB and SNLI. Entropy for human highlights was calculated based on non-neutral examples.

When $\beta = 0$ no information about the label predicted by the classifier is being exposed to the explainer, and as a result we have a model that resembles a combination of selective (during training) and top-k (during test time) sparsemax explainers. This means that the results between these explainers are expected to be very similar in terms of CSR.[17] Overall, for both datasets, we can see a tradeoff between CSR and entropy H as β increases, suggesting that CSR is not able to capture the notion of quality (which was expected due to the subjective nature of an explanation). For IMDB the entropy values were lower than our previous explainers, but for SNLI they were very similar. A potential reason for this is the particularity of the two datasets: IMDB have long documents (280 words on average) with a large set of repetitive words which are not stopwords and are strongly correlated with the labels (e.g. good, ok, bad, etc.); SNLI

[16] In practice, we simply set attention scores associated with stopwords to $-\infty$.

[17] Note that this also depends on the performance of C and C_{sp}, which are indeed very similar in this case: 95.64 and 95.44.

premises are very short (14 words on average) without a large set of repetitive words. Finally, due to this tradeoff, we selected $\beta = 0.2$ for all of our experiments since it induces a very high CSR with a reasonably good entropy.

G Machine Translation experiments

G.1 Data

To compare explainers on a more challenging task with large $|\mathcal{Y}|$, we ran an experiment on neural machine translation (NMT), adapting the JoeyNMT framework (Kreutzer et al., 2019). We used the EN→DE IWSLT 2017 dataset (Cettolo et al., 2017), with the standard splits (Table 5).

G.2 Classifier

We replicated the work of Peters et al. (2019) with the exception that we used raw words as input instead of byte-pair encodings. The implementation is based on Joey-NMT (Kreutzer et al., 2019). We employed beam search decoding with beam size of 5, achieving a BLEU score of 20.49, 21.12 and 20.75 for softmax (C), 1.5-entmax (C_{ent}) and sparsemax (C_{sp}), respectively. We refer to the work of Peters et al. (2019) for more training details. Table 15 shows the classifier hyperparameters.

HYPERPARAM.	VALUE
Word embeddings size	512
BiRNN hidden size	512
Attention scorer	(Bahdanau et al., 2015)
Batch size	32
Optimizer	Adam
Number of epochs	100
Early stopping patience	8
Learning rate	0.001
Decrease factor	0.5
ℓ_2 regularization	0
RNN type	LSTM
RNN layers	2
Dropout	0.3
Hidden dropout	0.3
Maximum output length	100
Beam size	5

Table 15: Classifier hyperparmeters for neural machine translation.

G.3 Communication

We consider the decision taken by the NMT system when generating the t^{th} target word (y), given the source sentence x and the previously generated words $y_{1:t-1}$. Note that in this example $\mathcal{Y}$ is the entire target vocabulary. The message is the concatenation of k source words (ranked by importance, without any word order information) with the prefix $y_{1:t-1}$. The layperson must predict the target word given this limited information (see Fig. 3).

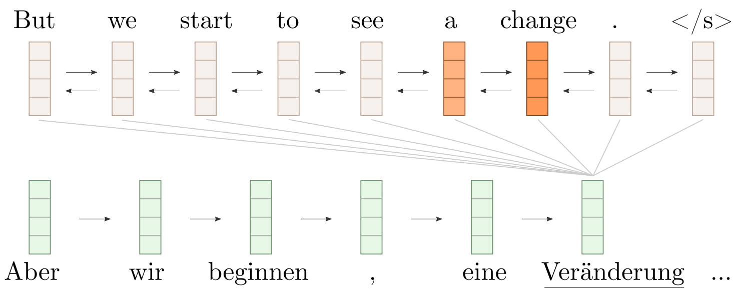


Figure 3: Example of sparse attention for machine translation. When the model is generating the word "Veränderung", the source words "a" and "change" are treated as explanation and sent as message.

The layperson is a model that uses an unidirectional LSTM with 256 hidden units to encode the translation prefix, and a feed-forward layer to encode the concatenation of k source word embeddings (the message) to a vector of 256 dimensions. The two vectors are concatenated and passed to a linear output layer to predict the next word $\tilde{y} \in \mathcal{Y}$ from the target vocabulary. We used 300D-pretrained GloVe embeddings to encode source words (EN), and 300D-pretrained FastText embeddings to encode target words (DE).[18] Table 16 shows the communication hyperparameters.

HYPERPARAM.	VALUE
Word embeddings size	300
LSTM hidden size	256
Merge LSTM states	concat
Batch size	16
Number of epochs	10
Early stopping patience	5
Learning rate	0.003
ℓ_2 regularization	10^{-5}

Table 16: Communication hyperparmeters for neural machine translation.

G.4 Results

Results comparing different filtering methods varying k are shown in Table 17. We show the CSR as we varied $k \in \{0, 1, 3, 5\}$. There are two main findings. First, we see again that **top-k attention outperforms top-k gradient**, in this case with a wider margin. Second, we see that all methods perform better as we increase k, albeit we can see a performance degradation of attention-based explainers for $k = 5$. An interesting case is when $k = 0$, meaning that L has no access to the source sentence, behaving like an unconditioned language model. In this case the performance is much worse, indicating that both explainers are selecting relevant tokens when $k > 0$. As we found for IMDB and SNLI, as we increased k we observed a trade-off between k and CSR for IWSLT. Fig. 4 depicts this finding.

CLF.	EXPLAINER	$k = 0$	$k = 1$	$k = 3$	$k = 5$
C	Top-k gradient	21.99	35.21	38.33	40.30
C	Top-k softmax	21.99	**62.58**	62.82	62.64
C_{ent}	Top-k 1.5-entmax	**22.31**	62.53	**63.48**	**62.69**
C_{sp}	Top-k sparsemax	22.14	62.21	61.94	61.92

Table 17: Results for IWSLT. Reported are CSR scores.

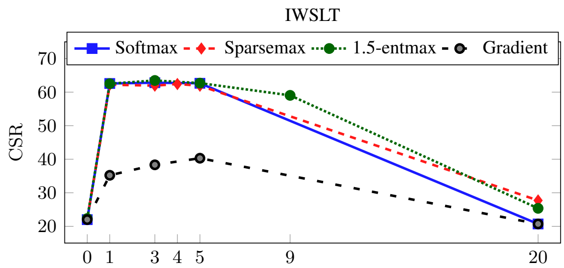
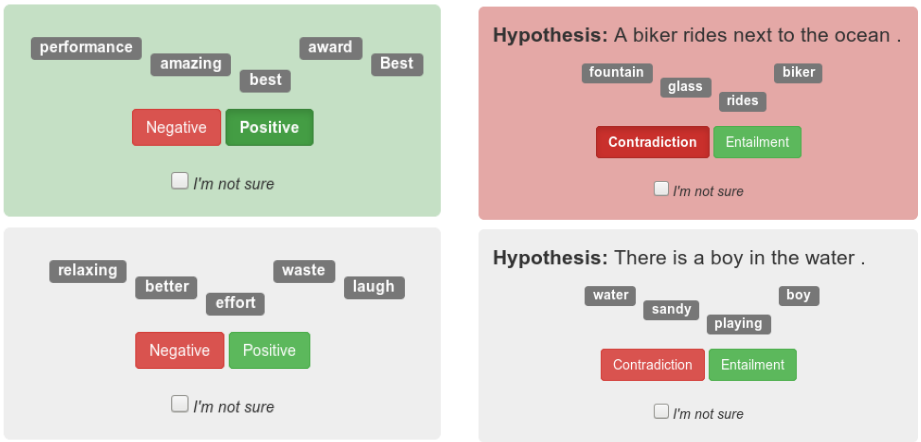


Figure 4: Message sparsity analysis for IWSLT. For SNLI, $k = 0$ corresponds to a case where the layperson only sees the translation prefix. The rightmost entry is the average length of the examples in the test set, and therefore it represents an explainer that simply pass forward all words to the layperson (i.e. a full bag-of-words). The average k for sparsemax and 1.5-entmax are, respectively: 4.5 and 9.4.

[18] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.de.300.bin.gz

H Human annotation

We had four different human annotators, two for IMDB and two for SNLI. No information was given about the explainers which produced each message, and documents were presented in random order. Since in our experiments we define the message as being a bag-of-words, which does not encode order information, the explanations (i.e. the selected words) were shuffled and displayed as a cloud of words. The annotators were asked to predict the label of each document, when seeing only these explanations. For SNLI, we show the entire hypothesis as raw text and the premise as a cloud of words. We selected top-k explainers with $k = 5$ for IMDB and $k = 4$ for SNLI. Figure 5 shows a snapshot of the annotation interface used for the experiments described in §6.



(a) IMDB (b) SNLI

Figure 5: Snapshot of the annotation interface.

By directly looking at the explanations, we observed that some of them are very ambiguous with respect to the true label, so we decided to include a checkbox to be marked in case the annotator was not sure by his/her decision. The unsure checkbox also helps to capture the notion of sufficiency, that is, if the explanations are sufficient for a human predict some label. A similar approach was employed by Yu et al. (2019) using a two-stage annotation method, explicitly asking the human annotator if the rationale was sufficient for his/her decision. Furthermore, we calculated the agreement between explainers using the Cohen's kappa coefficient and the relative observed agreement ratio (or accuracy, p_o). Table 18 shows statistics for the unsure checkbox and agreement between annotators.

CLF.	EXPLAINER	IMDB					SNLI				
		u	p_o	κ	CSR$_H$	ACC$_H$	u	p_o	κ	CSR$_H$	ACC$_H$
C	Erasure	0.05	0.92	0.83	89.25	86.25	0.25	0.83	0.66	72.50	83.50
C	Top-k gradient	0.17	0.76	0.51	73.50	73.00	0.32	0.80	0.59	65.75	76.75
C	Top-k softmax	0.23	0.91	0.81	89.25	88.25	0.25	0.78	0.55	72.00	82.75
C_{ent}	Top-k 1.5-entmax	0.09	0.91	0.81	89.25	85.75	0.29	0.82	0.64	70.00	80.50
C_{sp}	Top-k sparsemax	0.09	0.88	0.76	89.00	87.50	0.38	0.80	0.59	68.25	80.25
C_{ent}	Selec. 1.5-entmax	0.13	0.80	0.60	86.50	84.00	0.21	0.84	0.67	75.25	87.00
C_{sp}	Selec. sparsemax	0.10	0.89	0.77	87.75	86.75	0.35	0.83	0.66	72.25	85.00
C_{bern}	Bernoulli	0.25	0.72	0.43	79.00	75.00	0.24	0.85	0.69	74.50	86.75
C_{hk}	HardKuma	0.17	0.81	0.61	83.75	80.75	0.18	0.86	0.72	79.25	87.50
C	Joint E and L	0.12	0.96	0.91	96.75	89.25	0.65	0.71	0.44	58.00	70.00
-	Human highlights	-	-	-	-	-	0.34	0.88	0.74	83.25	83.25
Average		**0.14**	**0.85**	**0.70**	-	-	**0.31**	**0.82**	**0.63**	-	-

Table 18: Results for human evaluation. κ is the Cohen's kappa coefficient, p_o is the relative observed agreement, and u represents the average of the portion of examples where annotators were unsure about their decisions.

I Examples of explanations

Tables 19 and 20 show the average word overlap between explainers' messages (m) for IMDB and SNLI. Looking at the statistics we observed that, in general, top-k attention-based classifiers produce similar explanations among themselves, and the erasure explainer produces messages similar to top-k softmax. Major differences are observed for top-k gradient and rationalizers, while selective attention produces, by definition, more words than top-k attention (i.e. $m_{\text{top-k}} \subseteq m_{\text{selective}}$). It is worth noticing that although explainers with similar messages are expected to have a similar CSR (e.g. top-k attention and erasure), including/excluding a single word in the explanation might impact the layperson decision, as we can see in the next examples. Tables 21 and 22 show the output of erasure, gradient, attention, and joint explainers for IMDB, along with the prediction made by the classifier (y_C) and the layperson (y_L). In Tables 23 and 24 we also include the human highlights explainer for SNLI.

	Erasure	Top-k gradient	Top-k softmax	Top-k 1.5-entmax	Top-k sparsemax	Selec. 1.5-entmax	Selec. sparsemax	Bernoulli	HardKuma	Joint E and L
Erasure	*1.00*	0.34	0.85	0.56	0.55	0.20	0.37	0.14	0.23	0.20
Top-k gradient	0.34	*1.00*	0.35	0.30	0.30	0.16	0.26	0.11	0.18	0.11
Top-k softmax	0.85	0.35	*1.00*	0.57	0.55	0.20	0.37	0.14	0.24	0.20
Top-k 1.5-entmax	0.56	0.30	0.57	*1.00*	0.61	0.21	0.39	0.12	0.24	0.19
Top-k sparsemax	0.55	0.30	0.55	0.61	*1.00*	0.20	0.43	0.13	0.24	0.20
Selec. 1.5-entmax	0.20	0.16	0.20	0.21	0.20	*1.00*	0.45	0.24	0.44	0.08
Selec. sparsemax	0.37	0.26	0.37	0.39	0.43	0.45	*1.00*	0.21	0.41	0.13
Bernoulli	0.14	0.11	0.14	0.12	0.13	0.24	0.21	*1.00*	0.28	0.06
HardKuma	0.23	0.18	0.24	0.24	0.24	0.44	0.41	0.28	*1.00*	0.08
Joint E and L	0.20	0.11	0.20	0.19	0.20	0.08	0.13	0.06	0.08	*1.00*

Table 19: Average word overlap (%) between explainers for IMDB.

	Erasure	Top-k gradient	Top-k softmax	Top-k 1.5-entmax	Top-k sparsemax	Selec. 1.5-entmax	Selec. sparsemax	Bernoulli	HardKuma	Joint E and L
Erasure	*1.00*	0.38	0.77	0.55	0.41	0.35	0.37	0.32	0.49	0.38
Top-k gradient	0.38	*1.00*	0.40	0.36	0.31	0.34	0.33	0.32	0.35	0.26
Top-k softmax	0.77	0.40	*1.00*	0.56	0.41	0.36	0.37	0.32	0.49	0.38
Top-k 1.5-entmax	0.55	0.36	0.56	*1.00*	0.46	0.36	0.42	0.32	0.46	0.34
Top-k sparsemax	0.41	0.31	0.41	0.46	*1.00*	0.35	0.54	0.32	0.38	0.29
Selec. 1.5-entmax	0.36	0.34	0.36	0.36	0.35	*1.00*	0.64	0.88	0.48	0.26
Selec. sparsemax	0.37	0.33	0.37	0.42	0.54	0.64	*1.00*	0.60	0.45	0.26
Bernoulli	0.32	0.32	0.32	0.32	0.32	0.88	0.60	*1.00*	0.46	0.24
HardKuma	0.49	0.35	0.49	0.46	0.38	0.48	0.45	0.46	*1.00*	0.38
Joint E and L	0.38	0.26	0.38	0.34	0.29	0.26	0.26	0.24	0.38	*1.00*

Table 20: Average word overlap (%) between explainers for SNLI.

(**positive**) Mardi Gras : Made in china is an excellent movie that depicts how two cultures have much in common but , are not even aware of the influence each society has on one another . David Redmon open your eyes and allows you to see how the workers in china manufactures beads that cost little to nothing and are sold in America for up to 20 dollars . When Redmon questions Americans about where these beads come from they had no clue and seemed dumb founded . When he told them that they are made in China for less then nothing with horrible pay and unacceptable working conditions , Americans seemed sad , hurt , and a little remorseful but didn ' t really seem that they would stop purchasing the beads after finding out the truth . When Redmon questioned the workers in china they did not know that Americans were wearing them over their necks and paid so much for these beads . The workers laughed at what the purpose was behind beads and couldn ' t believe it . This movie is a great film that gives us something to think about in other countries besides our own . < br > < br > M . Pitts

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	pos	pos	excellent great film besides hurt
Top-k gradient	pos	neg	hurt horrible a excellent couldn
Top-k softmax	pos	pos	excellent great film movie besides
Top-k 1.5-entmax	pos	pos	great excellent couldn that besides
Top-k sparsemax	pos	pos	excellent great couldn gives besides
Select. entmax15	pos	pos	great excellent couldn that besides hurt didn that horrible is china Pitts gives us Redmon stop is not for t
Select. sparsemax	pos	pos	excellent great couldn gives besides china hurt that is
Bernoulli	pos	neg	an excellent movie another dumb horrible unacceptable sad remorseful movie great br br Pitts
HardKuma	pos	pos	excellent movie depicts America dumb horrible a great gives us besides our Pitts
Joint E and L	pos	pos	great excellent

(**negative**) I don ' t remember " Barnaby Jones " being no more than a very bland , standard detective show in which , as per any Quinn Martin show , Act I was the murder , Act II was the lead character figuring out the murder , Act III was the plot twist (another character murdered), Act IV was the resolution and the Epilogue was Betty (Lee Meriwether) asking her father - in - law Barnaby Jones (Buddy Ebsen) how he figured out the crime and then someone saying something witty at the end of the show . < br > < br > One thing I do remember was the late , great composer Jerry Goldsmith ' s excellent theme song . Strangely , the opening credit sequence made me want to see the show off and on for the seven seasons the show was on the air . I will also admit that it was nice to see Ebsen in a role other than Jed Clampett despite Ebsen being badly miscast . I just wished the show was more entertaining than when I first remembered it . < br > < br > Update (1 / 11 / 2009): I watched an interview with composer Jerry Goldsmith on YouTube through their Archive of American Television channel . Let ' s just say that I was more kind than Goldsmith about the show " Barnaby Jones ."

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neg	pos	wished excellent remembered miscast Strangely
Top-k gradient	neg	neg	miscast excellent remembered it badly
Top-k softmax	neg	pos	wished excellent remembered miscast figuring
Top-k 1.5-entmax	neg	neg	wished remembered Strangely miscast excellent
Top-k sparsemax	neg	neg	Strangely miscast wished badly excellent
Select. entmax15	neg	neg	wished remembered Strangely miscast excellent admit bland no character figuring say badly figured credit , the < the witty want just thing <
Select. sparsemax	neg	neg	Strangely miscast wished badly excellent remembered bland
Bernoulli	neg	neg	very bland , lead character plot character Epilogue witty show br late composer excellent theme song Strangely seasons nice badly miscast entertaining remembered br (1 / composer American Television
HardKuma	neg	neg	bland figuring saying excellent Strangely credit admit miscast wished remembered (1 11
Joint E and L	neg	neg	bland badly something

(**positive**) Yes ... I ' m going with the 1 - 0 on this and here ' s why . In the last few years , I have watched quite a few comedies and only left with a few mild laughs and a couple video rental late fees because the movies were that easy to forget . Then I stumble upon " Nothing ". Looked interesting , wasn ' t expecting much though . I was wrong . This was probably one of the funniest movies I have ever had the chance to watch . Dave and Andrew make a great comedic pair and the humor was catchy enough to remember , but not over complex to the point of missing the joke . I don ' t want to remark on any of the actual scenes , because I do feel this is a movie worth seeing for once . With more and more pointless concepts coming into movies (you know , like killer military jets and " fresh " remakes that are ruining old classics), This movie will make you happy to say it ' s OK to laugh at " Nothing ".

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	pos	pos	funniest worth great wrong pointless
Top-k gradient	pos	pos	comedic funniest OK worth joke
Top-k softmax	pos	pos	funniest worth great wrong pointless
Top-k 1.5-entmax	pos	pos	funniest great wrong worth not
Top-k sparsemax	pos	pos	funniest worth great catchy wrong
Select. entmax15	pos	neg	funniest great wrong worth not catchy do probably pointless easy feel ruining movie OK joke ever Yes seeing stumble comedic mild don wasn enough), forget because 0 for
Select. sparsemax	pos	neg	funniest worth great catchy wrong ruining 0 feel easy OK not pointless
Bernoulli	neg	neg	- few comedies few mild laughs couple movies stumble interesting wrong probably funniest movies Dave great comedic humor catchy joke scenes movie pointless movies fresh remakes ruining movie Nothing ".
HardKuma	neg	neg	0 stumble wrong probably one funniest great catchy not joke a movie worth seeing pointless ruining OK Nothing
Joint E and L	pos	neg	funniest pointless worth

(**negative**) I ' m not to keen on The Pallbearer , it ' s not too bad , but just very slow at the times . As the movie goes on , it gets a little more interesting , but nothing brilliant . I really like David Schwimmer and I think he ' s good here . I ' m not a massive Gwyneth Paltrow fan , but I don ' t mind her sometimes and she ' s okay here . The Pallbearer is not a highly recommended movie , but if you like the leads then you might enjoy it .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neg	pos	brilliant slow recommended nothing good
Top-k gradient	neg	pos	not nothing recommended slow brilliant
Top-k softmax	neg	pos	brilliant slow nothing recommended good
Top-k 1.5-entmax	neg	neg	slow brilliant nothing not recommended
Top-k sparsemax	pos	pos	slow brilliant nothing recommended good
Select. entmax15	neg	pos	slow brilliant nothing not recommended good enjoy highly very if you goes don okay , little it bad gets really
Select. sparsemax	pos	pos	slow brilliant nothing recommended good enjoy very bad highly
Bernoulli	neg	neg	Pallbearer , too bad slow times movie , brilliant good massive okay Pallbearer highly movie enjoy
HardKuma	neg	pos	slow nothing brilliant good okay highly recommended might enjoy
Joint E and L	neg	neg	nothing bad slow okay highly

Table 21: Examples of extracted explanations for IMDB.

(**positive**) Ok , when I rented this several years ago I had the worst expectations . Yes , the acting isn ' t great , and the picture itself looks dated , but as I sat there , a strange thing happened . I started to like it . The action is great and there are few scenes that make you jump . Brion James , maybe one of the greatest B - grade actors next to Bruce Campbell , is great as always . The story isn ' t bad either . Now I wouldn ' t rush out and buy it , but you won ' t waste your time at least watching this good b - grade post apocalyptic western .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	pos	pos	good great great grade waste
Top-k gradient	pos	neg	waste worst greatest grade t
Top-k softmax	pos	neg	good great great worst grade
Top-k 1.5-entmax	pos	pos	great waste great good greatest
Top-k sparsemax	pos	neg	great waste great good grade
Select. entmax15	pos	pos	great waste great good greatest great always Ok apocalyptic Yes make buy t grade isn worst but wouldn strange is
Select. sparsemax	pos	neg	great waste great good grade greatest your worst Yes Ok
Bernoulli	neg	neg	worst , acting , looks strange great scenes greatest actors great story bad , waste watching good apocalyptic western
HardKuma	pos	neg	worst great great always waste good apocalyptic
Joint E and L	pos	neg	great worst

(**negative**) I have read each and every one of Baroness Orczy ' s Scarlet Pimpernel books . Counting this one , I have seen 3 pimpernel movies . The one with Jane Seymour and Anthony Andrews i preferred greatly to this . It goes out of its way for violence and action , occasionally completely violating the spirit of the book . I don ' t expect movies to stick directly to plots , i gave up being that idealistic long ago , but if an excellent movie of a book has already been made , don ' t remake it with a tv movie that includes excellent actors and nice costumes , but a barely decent script . Sticking with the 80 ' s version .... Rahne

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neg	pos	excellent excellent script barely decent
Top-k gradient	neg	neg	barely decent script if but
Top-k softmax	neg	pos	excellent excellent script decent barely
Top-k 1.5-entmax	neg	pos	barely excellent excellent have Sticking
Top-k sparsemax	neg	pos	excellent excellent barely pimpernel decent
Select. entmax15	neg	pos	barely excellent excellent have Sticking preferred decent . It don to t script way if costumes Counting pimpernel Rahne , nice greatly t have
Select. sparsemax	neg	pos	excellent excellent barely pimpernel decent preferred nice t Sticking It
Bernoulli	pos	pos	Baroness Orczy pimpernel movies greatly occasionally movies plots excellent movie tv excellent actors nice costumes barely decent script ....
HardKuma	neg	pos	have pimpernel preferred way excellent excellent barely decent Sticking Rahne
Joint E and L	neg	neg	barely expect decent preferred completely

(**negative**) While I agree that this was the most horrendous movie ever made , I am proud to say I own a copy simply because myself and a bunch of my friends were extras (mostly in the dance club scenes , but a few others as well . This movie had potential with Bolo and the director of Enter the Dragon signed on , but as someone who was on set most every day I can tell you that Robert Clouse was an old and confused individual , at least during the making of this movie . It was a wonder he could find his way to the set everyday . I would also like to think that this might have been a better movie if a lot of it had not been destroyed in a fire at Morning Calm studios . I can ' t say that it would have been for sure , but it would be nice to think so . I was actually surprised that it was ever released , and that someone like Bolo would attach his name to it without a fight . Oh well . Also look at the extras for pro wrestler Scott Levy , AKA Raven . He was a wrestler in Portland at the time ... nice guy , very smart .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neg	pos	horrendous well well nice nice
Top-k gradient	neg	pos	well horrendous this well very
Top-k softmax	neg	pos	horrendous well Oh well nice
Top-k 1.5-entmax	neg	neg	horrendous Oh surprised had agree
Top-k sparsemax	neg	pos	horrendous smart nice Oh had
Select. entmax15	neg	pos	horrendous Oh surprised had agree nice smart others ever well ever but most nice movie proud like wonder . way few without . find but It making well actually be everyday
Select. sparsemax	neg	pos	horrendous smart nice Oh had ever ever few . wonder nice
Bernoulli	neg	pos	most horrendous bunch extras mostly scenes few This movie old movie everyday lot nice extras wrestler wrestler nice guy
HardKuma	neg	neg	horrendous bunch few confused wonder Oh guy very smart
Joint E and L	neg	neg	horrendous without

(**positive**) Having read some of the other comments here I was expecting something truly awful but was pleasantly surprised . REALITY CHECK : The original series wasn ' t that good . I think some people remember it with more affection than it deserved but apart from the car chases and Daisy Duke ' s legs the scripts were weak and poorly acted . The Duke boys were too intelligent and posh for backwood hicks , the shrunken Boss Hog was too cretinous to be evil and Rosco was just hyper throughout every screen moment . It ' s amazing the series actually lasted as long as it did because it ran out of story lines during the first series . < br > < br > Back to the movie . If you watch this film in it ' s own right , not as a direct comparison to however you remember the TV series , then it ' s not bad at all . The real star is of course the General Lee . The car chases and stunts are excellent and that ' s really what D . O . H . is all about . Johnny Knoxville is his usual eccentric self and along with Seann William Scott as Cousin Bo the pair make this film really funny in a hilarious Dumb - And - Dumber sort of way the TV series never achieved . The lovely Jessica Simpson is a natch as Miss Daisy , Burt Reynolds makes a much improved Boss Hog and M . C . Gainey makes a believably nasty Rosco P . Coltrane , the way he always should have been . < br > < br > If you don ' t like slapstick humour and crazy car stunts then you wouldn ' t be watching this film anyway because you should know what to expect . Otherwise if you want an entertaining car - action movie with a few good laughs that ' s not too taxing on the brain then go see this enjoyable romp with an open mind .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	pos	pos	horrendous well well nice nice
Top-k gradient	pos	neg	well horrendous this well very
Top-k softmax	pos	pos	horrendous well Oh well nice
Top-k 1.5-entmax	pos	pos	horrendous Oh surprised had agree
Top-k sparsemax	pos	neg	horrendous smart nice Oh had
Select. entmax15	pos	pos	horrendous Oh surprised had agree nice smart others ever well ever but most nice movie proud like wonder . way few without . find but It making well actually be everyday
Select. sparsemax	pos	pos	horrendous smart nice Oh had ever ever few . wonder nice
Bernoulli	pos	pos	most horrendous bunch extras mostly scenes few This movie old movie everyday lot nice extras wrestler wrestler nice guy
HardKuma	neg	neg	horrendous bunch few confused wonder Oh guy very smart
Joint E and L	pos	pos	horrendous without

Table 22: (continuation) Examples of extracted explanations for IMDB.

(**entailment**)
Premise: Children and adults swim in large pool with red staircase .
Hypothesis: A group of people are swimming .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	ent	ent	swim pool staircase adults
Top-k gradient	ent	con	adults pool swim large
Top-k softmax	ent	ent	swim pool large staircase
Top-k 1.5-entmax	ent	ent	swim pool large staircase
Top-k sparsemax	ent	ent	swim pool large adults
Select. entmax15	ent	ent	swim pool large staircase adults Children in and with
Select. sparsemax	ent	ent	swim pool large adults in
Bernoulli	ent	ent	Children and adults swim in large pool with red staircase .
HardKuma	ent	con	swim large pool staircase
Joint E and L	ent	con	pool swim staircase

(**contradiction**)
Premise: A group of Asian children are gathered around in a circle listening to an older male in a white shirt .
Hypothesis: A man is wearing a black shirt .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	con	ent	Asian white male children
Top-k gradient	con	ent	circle children gathered to
Top-k softmax	con	ent	Asian white male children
Top-k 1.5-entmax	con	con	white older a male
Top-k sparsemax	con	con	a male shirt Asian
Select. entmax15	con	con	white older a male Asian listening circle shirt of children around gathered a an group in in . A to are
Select. sparsemax	con	con	a male shirt Asian . an
Bernoulli	ent	ent	A group of Asian children are gathered around in a circle listening to an older male in a white shirt .
HardKuma	con	ent	group Asian male white shirt
Joint E and L	con	con	male group white

(**contradiction**)
Premise: A woman is pushing her bike with a baby carriage in front .
Hypothesis: A woman is pushing groceries in a cart .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	con	con	baby woman bike pushing
Top-k gradient	con	con	carriage bike her with
Top-k softmax	con	neu	baby woman carriage pushing
Top-k 1.5-entmax	con	con	carriage woman her baby
Top-k sparsemax	ent	con	baby carriage woman front
Select. entmax15	con	con	carriage woman her baby pushing front is A . a bike with
Select. sparsemax	ent	ent	baby carriage woman front pushing is
Bernoulli	con	con	A woman is pushing her bike with a baby carriage in front .
HardKuma	con	con	woman pushing bike carriage
Joint E and L	con	con	woman baby

(**neutral**)
Premise: A woman in a gray shirt working on papers at her desk .
Hypothesis: Lady working in her desk tensely to completed the task

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neu	neu	desk papers woman .
Top-k gradient	neu	neu	desk on shirt at
Top-k softmax	neu	neu	desk papers woman .
Top-k 1.5-entmax	neu	neu	desk papers working woman
Top-k sparsemax	neu	neu	desk papers woman working
Select. entmax15	neu	ent	desk papers working woman . on shirt her at in a
Select. sparsemax	neu	ent	desk papers woman working her A
Bernoulli	neu	neu	A woman in a gray shirt working on papers at her desk .
HardKuma	neu	neu	woman working papers at desk
Joint E and L	neu	neu	working desk woman papers

(**neutral**)
Premise: A brown dog with a blue muzzle is running on green grass .
Hypothesis: A mean dog is wearing a muzzle to keep it from attacking cats

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	neu	neu	dog brown running muzzle
Top-k gradient	neu	neu	with brown on green
Top-k softmax	neu	neu	dog brown running blue
Top-k 1.5-entmax	con	con	dog blue brown muzzle
Top-k sparsemax	neu	neu	dog muzzle with is
Select. entmax15	con	neu	dog blue brown muzzle running is . A grass green with on a
Select. sparsemax	neu	neu	dog muzzle with is A a running on brown
Bernoulli	neu	con	A brown dog with a blue muzzle is running on green grass .
HardKuma	neu	neu	dog muzzle running
Joint E and L	neu	neu	dog running muzzle

Table 23: Examples of extracted explanations for SNLI.

(**contradiction**)
Premise: A man sits at a table in a room .
Hypothesis: A woman sits .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	con	ent	sits table . at
Top-k gradient	con	ent	. sits table A
Top-k softmax	con	ent	sits table . room
Top-k 1.5-entmax	con	ent	table . sits man
Top-k sparsemax	con	ent	man sits A at
Select. entmax15	con	ent	table . sits man A room a a at in
Select. sparsemax	con	ent	man sits A at in a a
Bernoulli	con	ent	A man sits at a table in a room .
HardKuma	con	con	man sits at
Joint E and L	con	con	man
Human Highlights	con	ent	man

(**entailment**)
Premise: Elderly woman climbing up the stairs .
Hypothesis: The old lady was walking up the stairs .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	ent	ent	stairs woman Elderly climbing
Top-k gradient	ent	con	Elderly stairs . the
Top-k softmax	ent	con	stairs woman Elderly climbing
Top-k 1.5-entmax	ent	con	stairs Elderly woman climbing
Top-k sparsemax	ent	con	stairs Elderly woman climbing
Select. entmax15	ent	con	stairs Elderly woman climbing up . the
Select. sparsemax	ent	con	stairs Elderly woman climbing the
Bernoulli	con	con	Elderly woman climbing up the stairs .
HardKuma	ent	con	Elderly woman climbing up stairs
Joint E and L	ent	ent	stairs Elderly climbing woman
Human Highlights	ent	con	Elderly woman climbing

(**entailment**)
Premise: A woman with a blond ponytail and a white hat is riding a white horse , inside a fence with a horned cow .
Hypothesis: The woman is riding a horse .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	ent	con	horse riding . fence
Top-k gradient	ent	ent	cow horse fence riding
Top-k softmax	ent	ent	horse riding fence cow
Top-k 1.5-entmax	ent	con	horse riding woman a
Top-k sparsemax	ent	con	horse riding a is
Select. entmax15	ent	con	horse riding woman a cow fence horned a is ponytail , a with inside blond A . hat
Select. sparsemax	ent	con	horse riding a is with A ,
Bernoulli	ent	con	A woman with a blond ponytail and a white hat is riding a white horse , inside a fence with a horned cow .
HardKuma	ent	con	woman ponytail riding horse inside horned cow
Joint E and L	ent	ent	cow horse fence inside
Human Highlights	ent	ent	woman blond horse fence horned cow

(**contradiction**)
Premise: A woman in a black coat eats dinner while her dog looks on .
Hypothesis: A woman is wearing a blue coat .

EXPLAINER	y_C	y_L	EXPLANATION
Erasure	con	ent	coat black woman dog
Top-k gradient	con	ent	dog eats black looks
Top-k softmax	con	ent	coat black woman dinner
Top-k 1.5-entmax	con	con	black coat woman dog
Top-k sparsemax	con	con	black a woman A
Select. entmax15	con	con	black coat woman dog a looks in . dinner eats her A on while
Select. sparsemax	con	ent	black a woman A coat in her
Bernoulli	con	ent	A woman in a black coat eats dinner while her dog looks on .
HardKuma	con	ent	woman black coat
Joint E and L	con	con	woman black
Human Highlights	con	con	black

Table 24: (continuation) Examples of extracted explanations for SNLI.