Are Shortest Rationales the Best Explanations for Human Understanding?

Hua Shen[†]   **Tongshuang Wu**[◇]   **Wenbo Guo**[†]   **Ting-Hao 'Kenneth' Huang**[†]

[†]College of Information Sciences and Technology, Pennsylvania State University
[◇]Paul G. Allen School of Computer Science and Engineering, University of Washington
{huashen218,wzg13,txh710}@psu.edu
wtshuang@cs.washington.edu

Abstract

Existing self-explaining models typically favor extracting the shortest possible rationales — snippets of an input text "responsible for" corresponding output — to explain the model prediction, with the assumption that shorter rationales are more intuitive to humans. However, this assumption has yet to be validated. Is the shortest rationale indeed the most human-understandable? To answer this question, we design a self-explaining model, LIMITEDINK, which allows users to extract rationales at any target length. Compared to existing baselines, LIMITEDINK achieves compatible end-task performance and human-annotated rationale agreement, making it a suitable representation of the recent class of self-explaining models. We use LIMITEDINK to conduct a user study on the impact of rationale length, where we ask human judges to predict the sentiment label of documents based only on LIMITEDINK-generated rationales with different lengths. We show rationales that are too short do not help humans predict labels better than randomly masked text, suggesting the need for more careful design of the best human rationales.[1]

1 Introduction

While neural networks have recently led to large improvements in NLP, most of the models make predictions in a black-box manner, making them indecipherable and untrustworthy to human users. In an attempt to faithfully explain model decisions to humans, various work has looked into extracting *rationales* from text inputs (Jain et al., 2020; Paranjape et al., 2020), with *rationale* defined as the "shortest yet sufficient subset of input to predict the same label" (Lei et al., 2016; Bastings et al., 2019). The underlying assumption is two-fold: (1) by retaining the label, we are extracting the texts used by predictors (Jain et al., 2020); and (2) short
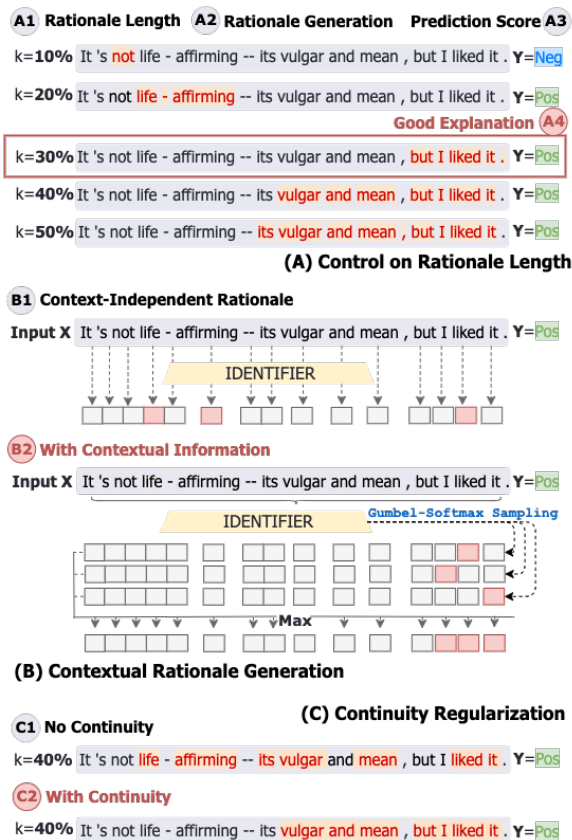

Figure 1: LIMITEDINK's rationale generation with length control: (A) control rationale generation with different lengths; (B) incorporating contextual information into rationale generation; (C) regularizing continuous rationale for human interpretability. Examples use the SST dataset for sentiment analysis (Socher et al., 2013).

rationales are more readable and intuitive for end-users, and thus preferred for human understanding (Vafa et al., 2021). Importantly, prior work has knowingly traded off some amount of model performance to achieve the shortest possible rationales. For example, when using less than 50% of text as rationales for predictions, Paranjape et al. (2020) achieved an accuracy of 84.0% (compared to 91.0% if using the full text). However, the assumption that the shortest rationales have better human interpretability has not been validated by

---

[1]Find open-source code at: https://github.com/huashen218/LimitedInk.git

human studies (Shen and Huang, 2021). Moreover, when the rationale is too short, the model has much higher chance of missing the main point in the full text. In Figure 1A, although the model can make the correct positive prediction when using only 20% of the text, it relies on a particular adjective, "life-affirming," which is seemingly positive but does not reflect the author's sentiment. These rationales may be confusing when presented to end-users.

In this work, we ask: *Are shortest rationales really the best for human understanding?* To answer the question, we first design LIMITEDINK, a self-explaining model that flexibly extracts rationales at any target length (Figure 1A). LIMITEDINK allows us to control and compare rationales of varying lengths on input documents. Besides **controls on rationale length**, we also design LIMITEDINK's sampling process and objective function to be **context-aware** (*i.e.,* rank words based on surrounding context rather than individually, Figure $1B_2$) and **coherent** (*i.e.,* prioritize continuous phrases over discrete tokens, Figure $1C_2$). Compared to existing baselines (*e.g.,* Sparse-IB), LIMITEDINK achieves compatible end-task performance and alignment with human annotations on the ERASER (DeYoung et al., 2020) benchmark, which means it can represent recent class of self-explaining models.

We use LIMITEDINK to conduct user studies to investigate the effect of rationale length on human understanding. Specifically, we ask MTurk participants to predict document sentiment polarities based on only LIMITEDINK-extracted rationales. By contrasting rationales at five different length levels, we find that shortest rationales are largely not the best for human understanding. In fact, humans do not perform better prediction accuracy and confidence better than using randomly masked texts when rationales are too short (*e.g.,* 10% of input texts). In summary, this work encourages a rethinking of self-explaining methods to find the right balance between brevity and sufficiency.

2 LIMITEDINK

2.1 Self-Explaining Model Definition

We start by describing typical self-explaining methods (Lei et al., 2016; Bastings et al., 2019; Paranjape et al., 2020). Consider a text classification dataset containing each document input as a tuple $(\mathbf{x}, y)$. Each input $\mathbf{x}$ includes n features (*e.g.,* sentences or tokens) as $\mathbf{x} = [x_1, x_2, ..., x_n]$, and y is the prediction. The model typically consists

of an *identifier* $\mathbf{idn}(\cdot)$ to derive a boolean mask $\mathbf{m} = [m_1, m_2, ..., m_n]$, where $m_i \in \{1, 0\}$ indicates whether feature x_i is in the rationale or not. Note that the mask $\mathbf{m}$ is typically a binary selection from the *identifier*'s probability distribution, i.e., $\mathbf{m} \sim \mathbf{idn}(\mathbf{x})$. Then it extracts rationales $\mathbf{z}$ by $\mathbf{z} = \mathbf{m} \odot \mathbf{x}$, and further leverages a *classifier* $\mathbf{cls}(\cdot)$ to make a prediction y based on the identified rationales as $y = \mathbf{cls}(\mathbf{z})$. The optimization objective is:

$$\min_{\theta_{\mathbf{idn}}, \theta_{\mathbf{cls}}} \underbrace{\mathbb{E}_{\mathbf{z} \sim \mathbf{idn}(\mathbf{x})} \mathcal{L}(\mathbf{cls}(\mathbf{z}), y)}_{\text{sufficient prediction}} + \underbrace{\lambda \Omega(\mathbf{m})}_{\text{regularization}} \quad (1)$$

where $\theta_{\mathbf{idn}}$ and $\theta_{\mathbf{cls}}$ are trainable parameters of *identifier* and *classifier*. $\Omega(\mathbf{m})$ is the regularization function on mask and λ is the hyperparameter.

2.2 Generating Length Controllable Rationales with Contextual Information

We next elaborate on the definition and method of controlling rationale length in LIMITEDINK Assuming that the rationale length is k as prior knowledge, we enforce the generated boolean mask to sum up to k as $k = \sum_{i=1}^{n}(m_i)$, where $\mathbf{m} = \mathbf{idn}(\mathbf{x}, k)$. Existing self-explaining methods commonly solve this by sampling from a Bernoulli distribution over input features, thus generating each mask element m_i independently conditioned on each input feature x_i (Paranjape et al., 2020). For example, in Figure $1B_1$), "life affirming" is selected independent of the negation context "not" before it, which contradicts with the author's intention. However, these methods potentially neglect the contextual input information. We leverage the concrete relaxation of subset sampling technique (Chen et al., 2018) to incorporate contextual information into rationale generation process (see Figure $1B_2$), where we aim to select the top-k important features over all n features in input $\mathbf{x}$ via Gumbel-Softmax Sampling (*i.e.,* applying the Gumbel-softmax trick to approximate weighted subset sampling process). To further guarantee precise rationale length control, we deploy the *vector and sort* regularization on mask $\mathbf{m}$ (Fong et al., 2019). See more model details in Appendix A.1.

2.3 Regularizing Rationale Continuity

To further enforce coherent rationale for human interpretability, we employ the Fused Lasso to encourage continuity property (Jain et al., 2020; Bastings et al., 2019). The final mask regularization is:

$$\Omega(\mathbf{m}) = \lambda_1 \underbrace{\sum_{i=1}^{n} |m_i - m_{i-1}|}_{\text{Continuity}} + \lambda_2 \underbrace{\| \text{vecsort}(m) - \hat{m} \|}_{\text{Length Control}} \quad (2)$$

Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER			
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	**.41**	.22	.83	.35	.52	.42
Sparse-IB	.84	.21	.42	.28	.46	**.17**	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	**.37**	.50	**.43**
LIMITEDINK	**.90**	**.26**	**.50**	**.34**	**.56**	.13	**.17**	**.15**	**.50**	**.04**	**.27**	**.07**	**.67**	**.22**	.40	**.28**	**.90**	.28	**.67**	.39
Length Level	50%				30%				50%				50%				40%			

Table 1: LIMITEDINK performs compatible with baselines in terms of end-task performance (**Task**, weighted average F1) and human annotated rationale agreement (**P**recision, **R**ecall, **F1**). All results are on test sets and are averaged across five random seeds. For LIMITEDINK, we report results for the best performing *length level*.

For BERT-based models, which use subword-based tokenization algorithms (*e.g.,* WordPiece), we assign each token's importance score as its subtokens' maximum score to extract rationales during model inference (see Figure 1*C*).

3 Model Performance Evaluation

We first validate LIMITEDINK on two common rationale evaluation metrics, including end-task performance and human annotation agreement.

3.1 Experimental Setup

We evaluate our model on five text classification datasets from the ERASER benchmark (DeYoung et al., 2020). We design the *identifier* module in LIMITEDINK as a BERT-based model, followed by two linear layers with the ReLU function and dropout technique. The temperature for Gumbel-softmax approximation is fixed at 0.1. Also, we define the *classifier* module as a BERT-based sequence classification model to predict labels. We train five individual self-explaining models of different rationale lengths with training and validation sets, where we set the rationale lengths as {10%, 20%, 30%, 40%, 50%} of all input text. Then we select one out of the five models, which has the best weighted average F1 score, to compare with current baselines on end-task performance and human annotation agreement on test sets. Note that we use all models with five rationale lengths in human evaluation described in Section 4.

Baselines. We compare LIMITEDINK with four baselines. `Full-Text` consists of only the *classifier* module with full-text inputs. `Sparse-N` enforces shortest rationales by minimizing rationale mask length (Lei et al., 2016; Bastings et al., 2019). `Sparse-C` controls rationale length by penalizing the mask when its length is less than a threshold (Jain et al., 2020). `Sparse-IB` enables length control by minimizing the KL-divergence between

the generated mask with a prior distribution (Paranjape et al., 2020). See Appendix A.1 for more model and baseline details.

3.2 Evaluation Results

End-Task Performance. Following metrics in DeYoung et al. (2020), we report the weighted average F1 scores for end-task classification performance. Among five LIMITEDINK models with different rationale lengths, Table 1 reports the model with the best end-task performance on the test set. We observe that LIMITEDINK performs similarly to or better than the self-explaining baselines in all five datasets. See ablation studies in Appendix A.2.

Human-Annotated Rationale Agreement. We calculate the alignment between generated rationales and human annotations collected in the ERASER benchmark (DeYoung et al., 2020). As also shown in Table 1, we report the Token-level F1 (F1) metric along with corresponding Precision (P) and Recall (R) scores. The results show that LIMITEDINK can generate rationales that are consistent with human annotations and comparable to self-explaining baselines in all datasets.

4 Human Evaluation

Equipped with LIMITEDINK, we next carry out human studies to investigate the effect of rationale length on human understanding.

4.1 Study Design

Our goal is to quantify human performance on predicting the labels and confidence based solely on the rationales with different lengths. To do so, we control LIMITEDINK to extract rationales of different lengths, and recruit Mechanical Turk (MTurk) workers to provide predictions and confidence.

Dataset & rationale extraction. We focus on sentiment analysis in user study, and randomly sample 100 reviews from the Movie Reviews (Zaidan

Part of Movie Review

"…….now he tries his hand at writing . …….. after you ' ve seen him in fargo and reservoir dogs , …. "

Q1: Is the movie review Positive or Negative?

Positive   Negative

Q2: How Confident are you in your above selection?

5-Very Confident 4-Pretty Confident 3-Hesitating 2-Not Confident 1-Random Guess

Figure 2: Key components of the User Interface in the MTurk *task* HITs. Note that each HIT contains five reviews with different rationale lengths.
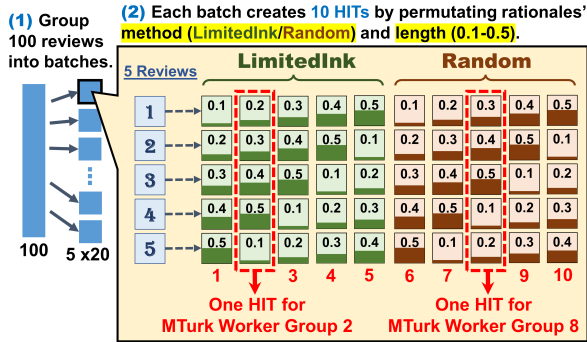

Figure 3: The human evaluation's workflow. We (1) divide 100 movie reviews into 20 batches and (2) produce 10 HITs from each batch for ten worker groups.

and Eisner, 2008) test set that have correct model predictions. Then, we extract five rationales for each review using **LIMITEDINK**, with lengths from 10% to 50%, with an increment of 10%.

Since human accuracy likely increases when participants see more words (*i.e.,* when the lengths of rationales increase), we also create a **Random** rationale baseline, where we randomly select words of the same rationale length on the same documents (10% to 50%) while taking the continuity constraint into consideration. More details of **Random** baseline generation are in Appendix A.3.1.

Study Procedure. The study is completed in two steps. First, we posted a *qualification* Human Intelligence Tasks (HITs, $0.50 per assignment) on MTurk to recruit 200 qualified workers.[2] Next, the 200 recruited workers can participate the *task* HIT ($0.20 per assignment, 7 assignments posted) which contains five distinct movie reviews, with varying rationale lengths (10%-50%). In *task* HIT, as key components shown in Figure 2, we only display the rationales and mask all other words with ellipses of random length, such that participants can not infer the actual review length. Then partic-

---

[2]In addition to our custom qualification used for worker grouping, three built-in worker qualifications are used in all of our HITs: HIT Approval Rate (≥98%), Number of Approved HITs (≥ 3000), and Locale (US Only) Qualification.
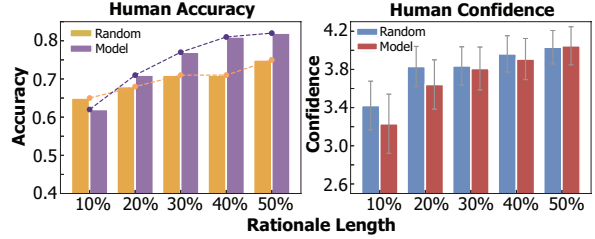

Figure 4: Human accuracy and confidence on predicting model labels given rationales with different lengths.

ipants are asked to guess the sentiment of the full review, and provide their confidence level based on a five-point Likert Scale (Likert, 1932). The full user interface is in Appendix A.3.2.

Participants recruiting and grouping. With each review having ten distinct rationales (five from LIMITEDINK and Random), if these rationale conditions were randomly assigned, participants are likely to see the same review repeatedly and gradually see all the words. We carefully design our study to eliminate such undesired learning effect. More specifically, we group our 100 reviews into 20 batches, with five reviews in each batch (Step 1 in Figure 3). For each batch, we create five HITs for LIMITEDINK and Random, respectively, such that all the rationale lengths of five reviews are covered by these 10 HITs (Step 2 in Figure 3). Further, we make sure each participant is only assigned to one unique HIT, so that each participant can only see a review once. To do so, we randomly divide the 200 qualified workers into 10 worker groups (20 workers per group), and pair one worker group with only one HIT in each batch. This way, each HIT can only be accomplished by one worker group. As our participant control is more strict than regular data labeling tasks on MTurk, we keep the HITs open for 6 days. 110 out of 200 distinct workers participated in the main study, and they completed 1,169 of 1,400 assignments.

4.2 Results

We show the human prediction accuracy and confidence results in Figure 4. We find that the best explanations for human understanding are largely not the shortest rationales (10% length level): here, the human accuracy in predicting model labels is lower than for the random baseline (0.61 vs. 0.63), indicating that the shortest rationales are not the best for human understanding. There is a significant difference in human predicted labels (*i.e.,* "positive"=1,"negative"=2) between LIMITEDINK (M=1.24,SD=0.71) and Random

13

length level (%) & Extract. method		Negative P / R / F1	Positive P / R / F1
10%	LIMITEDINK	0.66 / 0.56 / / 0.61	**0.70** / 0.58 / 0.64
	Random	**0.67 / 0.57 / 0.62**	0.66 / **0.70 / 0.68**
20%	LIMITEDINK	**0.75 / 0.61 / 0.67**	**0.71 / 0.77 / 0.74**
	Random	0.69 / 0.60 / 0.64	0.68 / 0.74 / 0.71
30%	LIMITEDINK	**0.74 / 0.76 / 0.75**	**0.81 / 0.78 / 0.79**
	Random	0.72 / 0.61 / 0.66	0.72 / 0.78 / 0.75
40%	LIMITEDINK	**0.84 / 0.76 / 0.80**	**0.78 / 0.85 / 0.81**
	Random	0.79 / 0.63 / 0.70	0.65 / 0.79 / 0.71
50%	LIMITEDINK	**0.78 / 0.78 / 0.78**	**0.85 / 0.84 / 0.85**
	Random	0.77 / 0.63 / 0.70	0.75 / 0.84 / 0.79

Table 2: Human performance (*i.e.,* Precision / Recall / F1 Score) on predicting model labels of each category in the Movie Reviews dataset.

(M=1.32,SD=0.54); t(1169)=2.27, p=0.02. Table 2 shows human performance for each category.

Additionally, notice that the slope of our model's accuracy consistently flattens as the rationale increases, whereas the random baseline does not display any apparent trend and is obviously lower than our model at higher length levels (*e.g.*, 40%). We hypothesize that this means our model is (1) indeed learning to reveal useful rationales (rather than just randomly displaying meaningless text), and (2) the amount of information necessary for human understanding only starts to saturate at around 40% of the full text. This creates a clear contrast with prior work, where most studies extract 10-30% of the text as the rationale on the same dataset (Jain et al., 2020; Paranjape et al., 2020). The eventually flattened slope potentially suggests a sweet spot to balance human understanding on rationales and sufficient model accuracy.

5 Discussion

By examining human prediction performance on five levels of rationale lengths, we demonstrate that the shortest rationales are largely not the best for human understanding. We are aware that this work has limitations. The findings are limited to Movie Reviews dataset, and we only evaluate human performance with rationales generated by the proposed LIMITEDINK. Still, our findings challenge the "shorter is better" assumption commonly adopted in existing self-explaining methods. As a result, we encourage future work to more cautiously define the best rationales for human understanding, and trade off between model accuracy and rationale length. More concretely, we consider that rationale models should find the right balance between

brevity and sufficiency. One promising direction could be to clearly define the optimal human interpretability in a measurable way and then learn to adaptively select rationales with appropriate length.

6 Related Work

Self-explaining models. Self-explaining models, which condition predictions on their rationales, are considered more trustworthy than post-hoc explanation techniques (Rajagopal et al., 2021). However, existing efforts often enforce minimal rationale length, which degrade the predictive performance (Yu et al., 2019; Bastings et al., 2019; Jain et al., 2020). Paranjape et al. (2020) improves this by proposing an information bottleneck approach to enable rationale length control at the sentence level. In this paper, LIMITEDINK further enables length control at the token level to allow more flexibility needed for our human studies.

Human-grounded evaluation. A line of studies evaluated model-generated rationales by comparing them against human-annotated explanations (Carton et al., 2020; Paranjape et al., 2020). Some other studies collect feedback from users to evaluate the explanations, such as asking people to choose a preferred model (Ribeiro et al., 2016) or to guess model predictions only based on rationales (Lertvittayakumjorn and Toni, 2019; Shen and Huang, 2020).

7 Conclusion

To investigate if the shortest rationales are best understandable for humans, this work presents a self-explaining model, LIMITEDINK, that achieves comparable performance with current self-explaining baselines in terms of end-task performance and human annotation agreement. We further use LIMITEDINK to generate rationales for human studies to examine how rationale length can affect human understanding. Our results show that the shortest rationales are largely not the best for human understanding. This would encourage a rethinking of rationale methods to find the right balance between brevity and sufficiency.

8 Acknowledgment

9 Ethical Considerations

This work shows that the shortest rationales are often not the best for human understanding. We thus advocate for studying how users interact with machine-generated rationales. However, we are aware that using rationales to interpret model prediction could pose some risks for users. Rationales omit a significant portion of the contents (in our case, 50% to 90% of the words in a movie review are omitted), which could convey information incorrectly or mislead users. Furthermore, machine-learned rationales could encode some unwanted biases (Chuang et al., 2021). We believe that such risks should be explicitly communicated with users in real-world applications.

References

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 882–891. PMLR.

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 114–120, Online. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2950–2958. IEEE.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,*

San Francisco, CA, USA, August 13-17, 2016, pages 1135–1144. ACM.

Hua Shen and Ting-Hao Huang. 2020. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172.

Hua Shen and Ting-Hao'Kenneth' Huang. 2021. Explaining the road not taken. *ACM CHI 2022 Workshop on Human-Centered Explainable AI*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. 2021. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

A Appendix

A.1 Model Details and Hyperparameters

A.1.1 Methodology Details

Concrete Relaxation of Subset Sampling Process. Given the output logits of *identifier*, we use Gumbel-softmax (Jang et al., 2017) to generate a concrete distribution as $\mathbf{c} = [c_1, ...c_n] \sim$ Concrete($\mathbf{idn}(\mathbf{x})$), represented as a one-hot vector over n features where the top important feature is 1. We then sample this process k times in order to sample top-k important features, where we obtain k concrete distributions as $\{\mathbf{c}^1, ..., \mathbf{c}^k\}$. Next we define one n-dimensional random vector $\mathbf{m}$ to be the element-wise maximum of these k concrete distributions along n features, denoted as $\mathbf{m} = \max_j \{\mathbf{c}_i^j\}_{i=n}^{j=k}$. Discarding the overlapping features to keep the rest, we then use $\mathbf{m}$ as the k-hop vector to approximately select the top-k important features over document $\mathbf{x}$.

Vector and sort regularization. We deploy a *vector and sort* regularization on mask $\mathbf{m}$ (Fong et al., 2019), where we sort the output mask m in a increasing order and minimize the L_1 norm between m and a reference $\hat{m}$ consisting of $n-k$ zeros followed by k ones.

A.1.2 Model Training Details

Training and inference. During training, we select the Adam optimizer with the learning rate at 2e-5 with no decay. We set hyperparameters in Equation 5 and 2 as $\lambda = 1e-4$, $v_1 = 0.5$ and $v_2 = 0.3$ and trained 6 epochs for all models. Furthermore, we train LIMITEDINK on a set of sparsity levels as $k = \{10\%, 20\%, 30\%, 40\%, 50\%\}$ and choose models with optimal predictive performance on validation sets.

A.1.3 Details of Self-Explaining Baselines

We compare our method with state-of-the-art self-explaining baseline models.

Sparse-N (Minimization Norm). This method learns the short mask with minimal L_0 or L_1 norm (Lei et al., 2016; Bastings et al., 2019), which penalizes for the total number of selected words in the explanation.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn}(\mathbf{x})} \mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda\|m\| \qquad (3)$$

Sparse-C (Controlled Norm Minimization). This method controls the mask sparsity through

a tunable predefined sparsity level α (Chang et al., 2020; Jain et al., 2020). The mask is penalized as below as long as the sparsity level α is passed.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn}(\mathbf{x})}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda \max(0, \frac{\|\mathbf{m}\|}{N} - \alpha)$$
(4)

where N is the input length and $\|m\|$ denotes mask penalty with L_1 norm.

Sparse IB (Controlled Sparsity with Information Bottleneck). This method introduces a prior probability of **z**, which approximates the marginal $p(\mathbf{m})$ of mask distribution; and $p(\mathbf{m}|\mathbf{x})$ is the parametric posterior distribution over **m** conditioned on input **x** (Paranjape et al., 2020). The sparsity control is achieved via the information loss term, which reduces the KL divergence between the posterior distribution $p(\mathbf{m}|\mathbf{x})$ that depends on **x** and a prior distribution $r(\mathbf{m})$ that is independent of **x**.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn}(\mathbf{x})}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda KL[p(\mathbf{m}|\mathbf{x}), r(\mathbf{m})]$$
(5)

A.2 Ablation Study on Model Components

We provide an ablation study on the Movie dataset to evaluate each loss term's influence on end-task prediction performance, including Precision, Recall, and F1 scores. The result is shown in Table 3.

Setups	End-Task Prediction		
	Precision	Recall	F1
No Sufficiency	0.25	0.50	0.34
No Continuity	0.82	0.81	0.81
No Sparsity	0.80	0.79	0.79
No Contextual	0.83	0.83	0.83
Our Model	**0.91**	**0.90**	**0.90**

Table 3: Ablation study of each module in our model on Movie Review dataset.

A.3 Additional Details of Human Study

A.3.1 Generating Random Baselines

Human accuracy likely increases when participants can see more words, *i.e.*, when the lengths of rationales increase. If a rationale and a random text span have the same number of words, the rationale should help readers predict the label better. We created a simple baseline that generated rationales by randomly selecting words to form the rationales.

We could control (1) how many words to select and (2) how many disjointed rationales to produce. In the study, we set these two numbers to be identical to that of LIMITEDINK at each length level.

In detail, given the rationale length k, we first got the count of total tokens in rationale as #tokens = k. Next, we computed the average number of rationale segments m, which are generated by LIMITEDINK, over the Movie dataset. We randomly selected m spans with total tokens' count as #tokens from the full input texts, thus obtaining the random baselines. We evenly separated 10 worker groups to finish five random baseline HITs and LIMITEDINK HITs each. We determined that good model rationales should get higher human accuracy compared with same-length random baselines.

A.3.2 Human Evaluation User Interface

We provide our designed user interfaces used in the human study. Specifically, we show the interface of the human study panel in Figure 5 (B). We also provide the detailed instructions for workers to understand our task, the instruction inteface is shown in Figure 6.

	Review1	Review2	Review3	Review4	Review5
Worker Group 1	Our@10%	Our@20%	Our@30%	Our@40%	Our@50%
Worker Group 2	Our@20%	Our@30%	Our@40%	Our@50%	Our@10%
Worker Group 3	Our@30%	Our@40%	Our@50%	Our@10%	Our@20%
Worker Group 4	Our@40%	Our@50%	Our@10%	Our@20%	Our@30%
Worker Group 5	Our@50%	Our@10%	Our@20%	Our@30%	Our@40%
Worker Group 6	Random@10%	Random@20%	Random@30%	Random@40%	Random@50%
Worker Group 7	Random@20%	Random@30%	Random@40%	Random@50%	Random@10%
Worker Group 8	Random@30%	Random@40%	Random@50%	Random@10%	Random@20%
Worker Group 9	Random@40%	Random@50%	Random@10%	Random@20%	Random@30%
Worker Group 10	Random@50%	Random@10%	Random@20%	Random@30%	Random@40%

(A) Worker Group Assignment

Instructions

In this HIT, you will see **parts of a movie review**. Read it carefully, and:

(1) Based on the partial content you see, try your best to **guess the original movie review is Positive** or **Negative** toward the movie (i.e., the Sentiment of the review), and

(2) Tell us how **confident** you are about the guess.

In this HIT, you will label **five** movie reviews 😊.

Examples (Click to Show Examples)

Select Sentiment and Confidence of the Displayed Parts of Moview Review

Please select the **sentiment label of the displayed parts of the movie review** and provide your **confidence on the selection**.

Parts of the Movie Review 1

··············· recall hearing species 2 described as " erotic . " i would love to know who used with that adjective for this ·············· a woman ' s abdomen as an alien baby claws its way free , splat blood and gore in all directions . anyone turned on by that

Question1: Is the movie review **Positive** or **Negative**? Please guess based on the parts of texts you see.

Positive   Negative

It's an Empty Input (Empty reviews are usually caused by data processing errors)

Question2: How **Confident** are you in your above selection?

5 - Very Confident - The displayed texts show clear attitude, and reflects the core sentiment (like/dislike) of the full review.

4 - Pretty Confident - The displayed texts show attitude towards the movie, but not very clear to reflect the core sentiment.

3 - Hesitating - The displayed texts seem positive/negative, but I cannot guess if it's representative of the full review.

2 - Not Confident - The displayed texts are ambiguous. I am not confident on the attitude towards the movie.

1 - I Guess Randomly - The displayed texts are too trivial and does not reflect on the larger themes.

Submit

(B) Worker Study Interface

Figure 5: (A) The design of the worker group assignment in our human study. (B) The worker interface of the human study.

Instructions

Examples (Click to Hide Examples)

Here is a movie review example, with a **Positive** sentiment label as ground truth:

" trees lounge is the directoral debut from one of my favorite actors , steve busce . he gave memorable performance in in the soup , fargo , and reservoir dogs . now he tries his hand at writing , directing and acting all in the same flick . the movie starts out awfully slow with tommy (busce) hanging around a local bar the " trees lounge " and him pestering his brother . it ' s obvious he a loser . but as he says " it ' s better i ' m a loser and know i am , then being a loser and not thinking i am . " well put . the story starts to take off when his uncle dies , and tommy , not having a job , decides to drive an ice cream truck . well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevi) and . . . i liked this movie alot even though it did not reach my expectation . after you ' ve seen him in fargo and reservoir dogs , you know he is capable of a better performance . i think his brother , michael , did an excellent job for his debut performance . mr . busce is off to a good career as a director ! "

In the HIT, we will **hide the sentiment label** and **highlight part of texts** in this movie review. Then you'll be asked to:

(1) **guess the review's sentiment label** given only highlighted content you see;

(2) tell us **your confidence** on the selection.

Here we provide examples explaining **several different confidence levels** for your reference.

Example-1:

" .... i liked this movie alot even though it did not reach my expectation . ... i think his brother , michael , did an excellent job for his debut performance . mr . busce is off to a good career as a director !"

You Selected Label: Positive

Confidence: 5 - Very Confident - The displayed texts show clear attitude, and reflects the core sentiment (like/dislike) of the full review.

Explanation: The displayed texts **clearly show the writer's sentimental opinion** on the movie, such as "i liked this movie alot". You could be **Very Confident** to select your sentiment label in this example.

Example-2:

" it ' s obvious he a loser . but as he says " it ' s better i ' m a loser and know i am , then being a loser and not thinking i am .... well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevi) and . ..."

You Selected Label: Positive

Confidence: 3 - Hesitating - The displayed texts seem positive/negative, but I cannot guess if it's representative of the full review.

Explanation: The displayed texts seem positive / negative, such as "finding a love interest in", "it ' s obvious he a loser ". **BUT they are describing movie plot but not direct evidence on showing writer's sentimental opinions** on this movie. You might be **Hesitating** to select your sentiment label in this example.

Example-3:

" ....now he tries his hand at writing . ... after you ' ve seen him in fargo and reservoir dogs ,..... "

You Selected Label: Negative

Confidence: 1 - I Guess Randomly - The displayed texts are too trivial and does not reflect on the larger themes.

Explanation: The displayed texts **don't show clear sentimental information** on this movie. You might randomly guess one label and choose **I Guess Randomly** as your confidenct.

Figure 6: User Interface of the instruction in the human study.

19