Learning to Faithfully Rationalize by Construction

Sarthak Jain
Khoury College of Computer Sciences
Northeastern University
jain.sar@northeastern.edu

Sarah Wiegreffe
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Byron C. Wallace
Khoury College of Computer Sciences
Northeastern University
b.wallace@northeastern.edu

Abstract

In many settings it is important for one to be able to understand *why* a model made a particular prediction. In NLP this often entails extracting snippets of an input text 'responsible for' corresponding model output; when such a snippet comprises tokens that indeed informed the model's prediction, it is a *faithful* explanation. In some settings, faithfulness may be critical to ensure transparency. Lei et al. (2016) proposed a model to produce faithful rationales for neural text classification by defining independent snippet extraction and prediction modules. However, the discrete selection over input tokens performed by this method complicates training, leading to high variance and requiring careful hyperparameter tuning. We propose a simpler variant of this approach that provides faithful explanations by construction. In our scheme, named FRESH, arbitrary feature importance scores (e.g., gradients from a trained model) are used to induce binary labels over token inputs, which an extractor can be trained to predict. An independent classifier module is then trained exclusively on snippets provided by the extractor; these snippets thus constitute faithful explanations, even if the classifier is arbitrarily complex. In both automatic and manual evaluations we find that variants of this simple framework yield predictive performance superior to 'end-to-end' approaches, while being more general and easier to train.[1]

1 Introduction

Neural models dominate NLP these days, but it remains difficult to know *why* such models make specific predictions for sequential text inputs. This problem has been exacerbated by the adoption of deep *contextualized* word representations, whose architectures permit arbitrary and interdependent interactions between all inputs, making it particularly difficult to know which inputs contributed to any specific prediction.

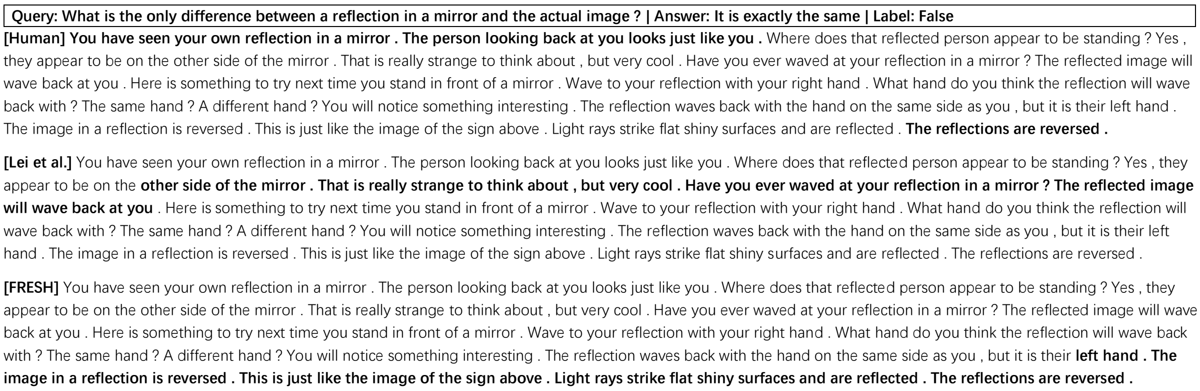
Concretely, in a bidirectional RNN or Transformer model, the *contextual embedding* for a word at position j in instance x may encode information from any or all of the tokens at positions 1 to j-1 and j+1 to $|x|$. Consequently, continuous scores such as attention weights (Bahdanau et al., 2015) induced over these contextualized embeddings reflect the importance not of individual inputs, but rather of unknown interactions between all input tokens. This makes it misleading to present heatmaps of these scores over the original token inputs as an explanation for a prediction (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019).

The key missing property here is *faithfulness* (Lipton, 2018): An explanation provided by a model is faithful if it reflects the information actually used by said model to come to a disposition. In some settings the ability of a model to provide faithful explanations may be paramount. For example, without faithful explanations, we cannot know whether a model is exploiting sensitive features such as gender (Pruthi et al., 2020).

We propose an approach to neural text classification that provides faithful explanations for predictions by construction. Following prior work in this direction (Lei et al., 2016), we decompose our model into independent extraction and prediction modules, such that the latter uses only inputs selected by the former. This discrete selection over inputs allows one to use an arbitrarily complex prediction network while still being able to guarantee that it uses only the extracted input features to inform its output.

The main drawback to this rationalization approach has been the difficulty of training the two components jointly under only instance-level

[1] Code is available at https://github.com/successar/FRESH



Query: What is the only difference between a reflection in a mirror and the actual image ? | Answer: It is exactly the same | Label: False

[Human] You have seen your own reflection in a mirror . The person looking back at you looks just like you . Where does that reflected person appear to be standing ? Yes , they appear to be on the other side of the mirror . That is really strange to think about , but very cool . Have you ever waved at your reflection in a mirror ? The reflected image will wave back at you . Here is something to try next time you stand in front of a mirror . Wave to your reflection with your right hand . What hand do you think the reflection will wave back with ? The same hand ? A different hand ? You will notice something interesting . The reflection waves back with the hand on the same side as you , but it is their left hand . The image in a reflection is reversed . This is just like the image of the sign above . Light rays strike flat shiny surfaces and are reflected . **The reflections are reversed .**

[Lei et al.] You have seen your own reflection in a mirror . The person looking back at you looks just like you . Where does that reflected person appear to be standing ? Yes , they appear to be on the **other side of the mirror . That is really strange to think about , but very cool . Have you ever waved at your reflection in a mirror ? The reflected image will wave back at you** . Here is something to try next time you stand in front of a mirror . Wave to your reflection with your right hand . What hand do you think the reflection will wave back with ? The same hand ? A different hand ? You will notice something interesting . The reflection waves back with the hand on the same side as you , but it is their left hand . The image in a reflection is reversed . This is just like the image of the sign above . Light rays strike flat shiny surfaces and are reflected . The reflections are reversed .

[FRESH] You have seen your own reflection in a mirror . The person looking back at you looks just like you . Where does that reflected person appear to be standing ? Yes , they appear to be on the other side of the mirror . That is really strange to think about , but very cool . Have you ever waved at your reflection in a mirror ? The reflected image will wave back at you . Here is something to try next time you stand in front of a mirror . Wave to your reflection with your right hand . What hand do you think the reflection will wave back with ? The same hand ? A different hand ? You will notice something interesting . The reflection waves back with the hand on the same side as you , but it is their **left hand . The image in a reflection is reversed . This is just like the image of the sign above . Light rays strike flat shiny surfaces and are reflected . The reflections are reversed .**

Figure 1: Contiguous rationales extracted using Lei et al. (2016) and FRESH models for an example from the MultiRC dataset. We also show the reference rationale associated with this example (top).

supervision (i.e., without token labels). This has necessitated training the extraction module via reinforcement learning — namely REINFORCE (Williams, 1992) — which exhibits high variance and is particularly sensitive to choice of hyperparameters. Recent work (Bastings et al., 2019) has proposed a differentiable mechanism to perform binary token selection, but this relies on the *reparameterization trick*, which similarly complicates training. Methods using the reparameterization trick tend to zero out token embeddings, which may adversely affect training in transformer-based models, especially when one is not fine-tuning lower layers of the model due to resource constraints, as in our experiments.

To avoid the complexity inherent to training under a remote supervision signal, we introduce **Faithful Rationale Extraction from Saliency tHresholding** (**FRESH**), which disconnects the training regimes of the extractor and predictor networks, allowing each to be trained separately. We still assume only instance-level supervision; the trick is to define a method of selecting snippets from inputs — *rationales* (Zaidan et al., 2007) — that can be used to support prediction. Here we propose using arbitrary feature importance scoring techniques to do so. Notably, these need not satisfy the 'faithfulness' criterion.

In this paper we evaluate variants of FRESH that use attention (Bahdanau et al., 2015) and gradient methods (Li et al., 2016; Simonyan et al., 2014) as illustrative feature scoring mechanisms. These provide continuous scores for features; we derive discrete rationales from them using simple heuristics. An independent network then uses *only* the extracted rationales to make predictions.

Disconnecting the training tie between the independent rationale extractor and prediction modules means that FRESH is faithful by construction: The snippet that is ultimately used to inform a prediction can be presented as a faithful explanation because this was the only text available to the predictor. In contrast to prior discrete rationalization methods, FRESH greatly simplifies training, and can accommodate any feature importance scoring metric. In our experiments, we also find that it yields superior predictive performance.

In addition to being faithful (and affording strong predictive performance), extracted rationales would ideally be intuitive to humans, i.e., *plausible*. To evaluate this we run a small user study (section 8) in which humans both evaluate the readability of extracted rationales and attempt to classify instances based on them, effectively serving as a prediction module in the FRESH framework. An example illustrating this property is presented in Figure 1.

2 Related Work

Types of explainability. Lipton (2018); Doshi-Velez and Kim (2017) and Rudin (2019) provide overviews on definitions and characterizations of interpretability. Lertvittayakumjorn and Toni (2019) classify three possible uses of text explanations: (*i*) revealing model behavior, (*ii*) justifying model predictions, and (*iii*) helping humans investigate uncertain predictions. Attempting to guarantee the faithfulness of a feature selection or explanation generation method is a more challenging question than finding explanations which humans find acceptable (Rudin, 2019). But the benefits of developing such methods is profound: Faithful explanations provide a means to reveal a

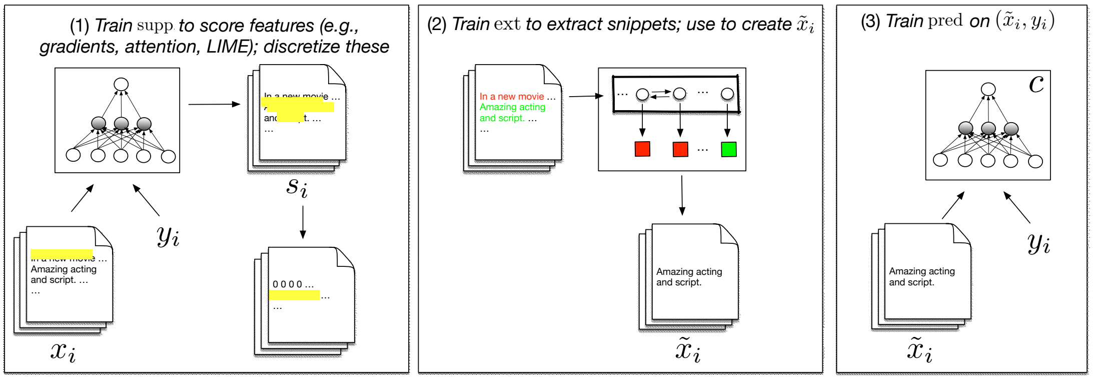


Figure 2: A schematic of the FRESH approach. (1) The first model, supp, is trained end-to-end for prediction but used only to ‘importance score’ features. These scores can be derived via any method, e.g., gradients or attention, and are not required to faithfully explain model outputs. Scores are heuristically discretized into binary labels. (2) An extraction module ext may be a parameterized sequence tagging model trained on the pseudo-targets derived in (1), or heuristics over importance scores directly, creating a new dataset $\langle \tilde{x}, y \rangle$ comprising pairs of extracted rationales only. (3) This new dataset is used to train a final classifier, pred, which only ever sees rationales.

model’s underlying decision-making process.

Issues with current explainability methods in NLP. A recent line of work in NLP has begun to critically examine the use of certain methods for constructing ‘heatmaps’ over input tokens to explain predictions. In particular, existing feature attribution methods may not provide robust, faithful explanations (Feng et al., 2018; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Brunner et al., 2020; Zhong et al., 2019; Pruthi et al., 2020).

Wiegreffe and Pinter (2019) argue for classifying model interpretability into two groups: faithfulness and plausibility. Lei et al. (2016) note that a desirable set of criteria for rationales is that they are sufficient, short, and coherent. Yu et al. (2019) extend these criteria by additionally arguing for comprehensiveness, which dictates that a rationale should contain all relevant and useful information.

Prior efforts (Lei et al., 2016; Yu et al., 2019; Bastings et al., 2019) have proposed methods that produce faithful explanations via a two-model setup, defining a *generator* network that imposes *hard attention* over inputs and then passes these to a second model for prediction. Yu et al. (2019) extend this by adding a third adversarial model into the framework. These models are trained jointly, which is difficult because hard attention is discrete and necessitates recourse to reinforcement learning, i.e., REINFORCE (Williams, 1992), or the reparameterization trick (Bastings et al., 2019).

Human evaluations. Kim et al. 2016 states: “a method is interpretable if a user can correctly predict the method’s result”; they conducted user studies to test this. In a similar plausibility vein, others have proposed testing whether humans *like* rationales (Ehsan et al., 2018, 2019). We follow these efforts by eliciting human judgments on rationales, although we view plausibility as a secondary aim here.

3 Faithfulness through Discrete Rationale Selection

We now propose FRESH, our framework for training explainable neural predictors. We begin by describing the two-model, discrete rationale selection approach introduced by Lei et al. (2016) (§3.1), which serves as the starting point for our framework, detailed in §4.

3.1 End-to-End Rationale Extraction

Consider a standard text classification setup in which we have n input documents $X = \{x_1, ..., x_n\}$, $x_i \in V^{l_i}$, where l_i denotes the number of tokens in document x_i, and V the vocabulary, and their assigned labels $y = \{y_1, ..., y_n\}$, $y_i \in \mathcal{Y}$. Lei et al. propose a model comprising a generator (gen) and an encoder (enc). gen is tasked with extracting rationales from inputs x_i, formalized as a binary mask over tokens sampled from a Bernoulli distribution: $z_i \sim \text{gen}(x_i) \in \{0, 1\}^{l_i}$. enc makes predictions $\hat{y} = \text{enc}(x_i, z_i)$ on the basis of the unmasked tokens.

The objective function is defined so that the overall expected loss $\mathcal{L}$ is minimized over both modules:

$$\underset{\theta_{\text{enc}},\theta_{\text{gen}}}{\text{minimize}} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim \text{gen}(x_i)} \mathcal{L}\left(\text{enc}(x_i, z_i), y_i\right). \quad (1)$$

This objective (1) is difficult to optimize as it requires marginalizing over all possible rationales z. Parameter estimation is therefore performed via an approximation approach that entails drawing samples from $\text{gen}(x)$ and averaging their associated gradients during the learning process. Lei et al. (2016) found that this REINFORCE-style estimation works well for rationale extraction, but may have high variance as a result of the large state space of possible rationales under consideration, which is difficult to efficiently explore.

The loss function $\mathcal{L}$ used by Lei et al. (2016) is a squared ℓ_2 loss between the prediction $\text{enc}(x, z)$ and the reference label y, with added regularization terms placed on the binary mask z to encourage rationale conciseness and contiguity.

We modify the conciseness term so that the model is not penalized as long as a predefined desired rationale length d has not been passed:

$$\Omega(z) = \underbrace{\lambda_1 \max\left(0, \frac{|z|}{L} - d\right)}_{\text{conciseness}} + \underbrace{\lambda_2 \sum_t \frac{|z_t - z_{t-1}|}{L-1}}_{\text{contiguity}}. \quad (2)$$

4 Faithful Rationale Extraction from Saliency tHresholding (FRESH)

To avoid recourse to REINFORCE, we introduce FRESH, in which we decompose the original prediction task into three sub-components, each with its own independent model. These are the **support** model supp, the rationale **extractor** model ext, and the **classifier** pred.[2]

We train supp end-to-end to predict y, using its outputs only to extract continuous feature importance scores from instances in X. These scores are binarized by ext either using a parameterized model trained on the output scores, or via direct discretization heuristics. Finally, pred is trained (and tested) *only on text provided by* ext. Figure 2 depicts this proposed framework.

A central advantage of our decomposed setup lies in the arbitrariness of the rationale extraction mechanism. Any function over supp's predictions that assigns scores to the input tokens intended to quantify their importance can serve as an input to ext. Note that this means even post-hoc scoring models (applied after the model has completed training) are permissible. Examples of such functions include gradient-based methods and LIME (Ribeiro et al., 2016).

Notably, the importance scoring function need not faithfully identify features that actually informed the predictions from supp. This means, e.g., that one is free to use token-level attention (over contextualized representations) — the final rationales provided by FRESH will nonetheless remain faithful with respect to pred. The importance scores are used only to train ext heuristically, for example by treating the top k tokens (with respect to importance scores) for a given example as the target rationale. The key design decision here is designing such heuristics that map continuous importance scores to discrete rationales. Any strategy for this will likely involve trading conciseness (shorter rationales) against performance (greater predictive accuracy).

For explainability, we can present users with the snippet(s) that pred used to make a prediction as an explanation (from ext), and we can be certain that the only tokens that contributed to the prediction made by pred are those included in the this text. In addition to transparency, this framework may afford efficiency gains in settings in which humans are tasked with classifying documents; in this case we can use ext to present only the (short) relevant snippets. Indeed, we use exactly this approach as one means of evaluation in Section 8.

5 FRESH Implementations

The high-level framework described above requires making several design choices to operationalize; we propose and evaluate a set of such choices in this work, detailed below. Specifically, we must specify a feature importance scoring mechanism for supp (Section 5.1), and a strategy for inducing discrete targets from these continuous scores (5.2). In addition, we need to specify a trained or heuristic extractor architecture ext. In this work, all instances of pred exploit BERT-based representations.[3]

[2] This is the most general framing, but in fact supp and ext may be combined by effectively defining ext as an application of heuristics to extract snippets on the basis of scores provided by supp; any means of procuring 'importance' scores for the features comprising instances and converting these to extracted snippets to pass to pred will suffice.

[3] For fair comparison, we have modified all baselines (Lei et al., 2016; Bastings et al., 2019) to similarly capitalize on BERT-based representations.

5.1 Feature Scoring Methods

All models considered in this work are based on Bidirectional Encoder Representations from Transformer (BERT) encoders (Devlin et al., 2019) and its variants, namely RoBERTa (Liu et al., 2019) and SciBERT (Beltagy et al., 2019); see Appendix B for more details. For sake of brevity, we simply refer to all of these as BERT from here on. We define supp as a BERT encoder that consumes either a single input (in the case of standard classification) or two inputs (e.g., in the case of question answering tasks) separated by the standard `[SEP]` token.

While we emphasize that the proposed framework can accommodate arbitrary input feature scoring mechanisms, we consider only a few obvious variants here, leaving additional exploration for future work. Specifically, we evaluate attention scores (Bahdanau et al., 2015) and input gradients (Li et al., 2016; Simonyan et al., 2014).

Attention scores are taken as the self-attention weights induced from the `[CLS]` token index to all other indices in the penultimate layer of supp; this excludes weights associated with any special tokens added. BERT uses wordpiece tokenization; to compute a score for a token, we sum the self-attention weights assigned to its constituent pieces. BERT is also multi-headed, and so we average scores over heads to derive a final score.

5.2 Discretizing Soft Scores

A necessary step in our framework consists of mapping from the continuous feature scores provided by supp to discrete labels, or equivalently, mapping scores to rationales which will either be consumed directly by pred or be used to train a sequence tagging model ext. We consider a few heuristic strategies for performing this mapping.

Contiguous. Select the span of length k that corresponds to the highest total score (over all spans of length k). We call these rationales **contiguous**.

Top-k. Extract as a rationale the top-k tokens (with respect to importance scores) from a document, irrespective of contiguity (each word is treated independently). We refer to these rationales as **non-contiguous**.

These strategies may be executed **per-instance** or **globally** (across an entire dataset), reflecting the flexibility of FRESH. Empirically, per-instance and global approaches performed about the same; we report results for the simpler, per-instance approaches (additional results in Appendix E).

	Doc. Len.	Rationale Len.	N
SST	17	-	9,613
AGNews	30	-	127,600
Ev. Inf.	349	10%	7,193
Movies	728	31%	1,999
MultiRC	297	18%	32,091

Table 1: Dataset details, with rationale length ratios included for datasets where they are available.

5.3 Extractor model

We experiment with two variants of ext. The first is simply direct use of the importance scores provided by supp and discretization heuristics over these; this does not require training an explicit ext model. We also consider a parameterized extractor model that independently makes token-wise predictions from BERT representations. Using an explicit extraction model allows us to mix in direct supervision on rationales alongside the pseudo-targets derived heuristically from supp.

Tying the sequential token predictions made by ext via a Conditional Random Field (CRF) layer (Lafferty et al., 2001) may further improve performance, but we leave this for future work.

6 Experimental Setup

6.1 Datasets

We use five **English** text classification datasets spanning a range of domains (see Table 1).

Stanford Sentiment Treebank (SST) (Socher et al., 2013). Sentences labeled with binary sentiment (neutral sentences have been removed).

AgNews (Del Corso et al., 2005). News articles to be categorized topically into *Science*, *Sports*, *Business*, and *World*.

Evidence Inference (Lehman et al., 2019). Biomedical articles describing randomized controlled trials. The task is to infer the reported relationship between a given intervention and comparator with respect to an outcome, and to identify a snippet within the text that supports this. The original dataset comprises lengthy full-text articles; we use an abstract-only subset of this data.

Movies (Zaidan and Eisner, 2008). Movie reviews labeled for sentiment accompanied by rationales on dev and test sets (DeYoung et al., 2020).

Saliency	Rationale	SST (20%)	AGNews (20%)	Ev. Inf. (10%)	Movies (30%)	MultiRC (20%)
Full text	–	.90 (.89-.90)	.94 (.94-.94)	.73 (.73-.78)	.95 (.93-.97)	.68 (.68-.69)
Lei *et al.*	contiguous	.71 (.49-.83)	.87 (.85-.89)	.53 (.45-.56)	.83 (.80-.92)	.62 (.62-.64)
	top k	.74 (.47-.84)	**.92 (.90-.92)**	.47 (.38-.53)	.87 (.80-.91)	.64 (.61-.65)
Bastings *et al.*	contiguous	.60 (.58-.62)	.77 (.18-.78)	.45 (.40-.49)	—	.41 (.30-.50)
	top k	.59 (.58-.61)	.72 (.19-.80)	.50 (.38-.60)	—	.44 (.30-.55)
Gradient	contiguous	.70 (.69-.72)	.85 (.84-.85)	.67 (.62-.68)	**.94 (.92-.95)**	**.67 (.66-.67)**
	top k	.68 (.67-.70)	.86 (.85-.86)	.62 (.61-.64)	.93 (.92-.94)	.66 (.65-.67)
[CLS] Attn	contiguous	**.81 (.80-.82)**	.88 (.88-.89)	**.68 (.59-.73)**	.93 (.90-.94)	.63 (.60-.62)
	top k	**.81 (.80-.82)**	**.91 (.90-.91)**	.66 (.64-.70)	**.94 (.93-.95)**	.63 (.62-.64)

Table 2: Model predictive performances across datasets, with rationale length as a percentage of each document in parentheses. We report mean Macro F1 scores on test sets, and min/max across random seeds. The top row (*Full text*) corresponds to a black-box model that does not provide explanations and uses the entire document; this is upper-bound on performance. We bold the best-performing rationalized model(s) for each corpus.

MultiRC (Khashabi et al., 2018). Passages and questions associated with multiple correct answers. Following DeYoung et al. (2020), we convert this to a binary classification task where the aim is to categorize answers as *True* or *False* based on a supporting rationale.

6.2 Model and Training Details

For datasets where human rationale annotations are available, we set k to the average human rationale annotation length, rounded to the nearest ten percent. For the rest, we set $k = 20\%$.

For generality, all models considered may consume both *queries* and texts, as is required for MultiRC and Evidence Inference. Rationales can be extracted from only from the text; this typically dominates the query in length, and is more informative in general. Further implementation details (including hyperparameters) are provided in Appendix A.

Hyperparameter sensitivity and variance. To achieve conciseness and contiguity, Lei et al. (2016) impose a regularizer on the encoder that comprises two terms (Equation 2) with associated hyperparameters (λ_1, λ_2). In practice, we have found that one needs to perform somewhat extensive hyperparameter search for this model to realize good performance. This is inefficient both in the sense of being time-consuming, and in terms of energy (Strubell et al., 2019).

By contrast, FRESH requires specifying and training independent module components, which incurs some energy cost. But there are no additional hyperparameters, and so FRESH does not require extensive hyperparameter search, which is typically the most energy-intensive aspect of model training. We quantify this advantage by reporting the variances over different hyperparameters we observed for (Lei et al., 2016) and the compute time this required to conduct this search in Appendix B.

In addition to being sensitive to hyperparameters, a drawback of REINFORCE-style training is that it can exhibit high variance within a given hyperparameter setting. To demonstrate this, we report the variance in performance of our proposed approach and of Lei et al. (2016) as observed over five different random seeds.

We also find that both Lei et al. (2016) and Bastings et al. (2019) tend to degenerate and predict either complete or empty text as rationale. To make results comparable to FRESH, at inference time, we restrict the rationale to specified desired length k before passing it to the corresponding classifier.

7 Quantitative Evaluation

We first evaluate the performance achieved on datasets by the pred models trained on different ext-extracted rationales, compared to each other and to Lei et al. (2016)'s end-to-end rationale extraction framework. As an additional baseline, we also evaluate a variant of the differentiable binary variable model proposed in Bastings et al. 2019. This baseline do not require any hyperparameter search.

In general, we would expect predictive performance to positively correlate with rationale length, and so we evaluate predictive performance (accuracy or F1-score) across methods using a fixed rationale length for each dataset.

We report results in terms of predictive performance for all model variants in Table 2. Here we

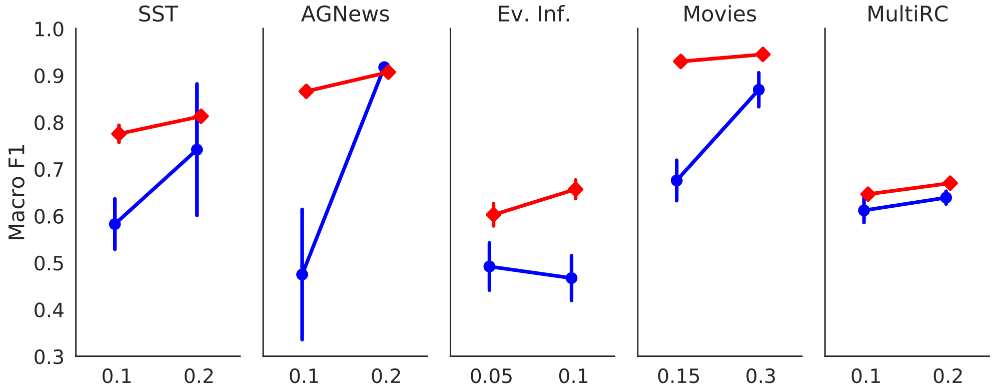


Figure 3: Results for Lei *et al.* (●) and FRESH (♦) evaluated across five datasets at two different desired rationale lengths (as % of document length). Vertical bars depict standard deviations observed over five random seeds.

use the entire train sets for the respective datasets, and fix the rationale length as described in §6.2 to ensure fair comparison across methods. We observe that despite its simplicity, FRESH performs nearly as well as *Full text* while using only 10-30% of the original input text, thereby providing transparency. FRESH achieves better average performance than Lei et al.'s end-to-end method, with the exception of AGNews, in which case the models are comparable. It also consistently fares better than Bastings et al.'s system.

Of the two feature scoring functions considered, `[CLS]` self-attention scores tend to yield better results, save for on the MultiRC and Movies datasets, on which gradients fare better. With respect to discretizing feature scores, the simple top-k strategy seems to perform a bit better than the contiguous heuristic, in what we expect to be traded off against a greater coherence of the contiguous rationales.

As seen in Table 2, FRESH exhibits lower variance across runs, and does not require hyperparameter search (further analysis in Appendix B).

Varying rationale length. Figure 3 plots F1 scores across datasets and associated standard deviations achieved by the best rationale variant of Lei et al. (2016) and FRESH at two different target rationale lengths. These results demonstrate the effectiveness of FRESH even in constrained settings. Note, we had to re-perform hyperparameter search for a different rationale length in case of (Lei et al., 2016) model.

Incorporating human rationale supervision. In some settings it may be feasible to elicit direct supervision on rationales, at least for a subset of training examples. Prior work has exploited such signal during training (Zhang et al., 2016; Strout et al., 2019; Small et al., 2011). One of the potential advantages of explicitly training the extraction model ext with pseudo-labels for tokens (derived from heuristics over importance scores) is the ability to mix in direct supervision on rationales alongside these derived targets.

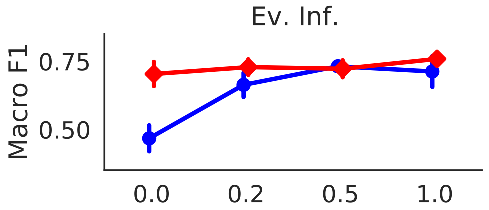


Figure 4: Results on Evidence Inference for Lei *et al.* (●) and FRESH (♦) given varying amounts of explicit rationale supervision.

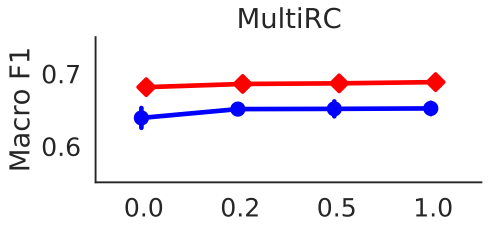


Figure 5: Results on MultiRC for Lei *et al.* (●) and FRESH (♦) given varying amounts of explicit rationale supervision.

We evaluate whether direct rationale supervision improves performance on two datasets for which we have human rationale annotations (Evidence Inference and MultiRC). In both cases we provide models with varying amounts of rationale-level supervision (0%, 20%, 50% and 100%), and again compare the best variants of Lei et al. (2016)

and our model. For the former, we introduce an additional binary cross entropy term into the objective for that explicitly penalizes the extractor for disagreeing with human token labels.

Explicitly training a sequence tagging model as ext over heuristic targets from supp did not improve results in our experiments. However, as shown in Figure 4 and Figure 5, mixing in rationale-level supervision when training ext *did* improve performance on the Evidence Inference dataset by a small amount, although not for MultiRC. This suggests that explicit rationale supervision may at least sometimes improve performance, and this is not possible without a parameterized ext model.

In Lei et al. (2016)'s framework, direct supervision provides considerable performance improvement in the case of Evidence Inference (although still suffering from variance effects), and did not affect performance on MultiRC.

8 Human Analysis

We have proposed FRESH as an architecture which, in addition to exceeding performance of previous training regimes, provides a guarantee for extracting rationales which are *faithful*. However, as noted in the introduction, another desirable trait of rationales is that they are judged as *good* by humans. To assess the plausibility of the resulting rationales (Herman, 2017; Wiegreffe and Pinter, 2019), we design a human user study.[4] We evaluate the following attributes of plausibility:

Sufficiency. Can a human predict the correct label given only the rationale? This condition aligns with Kim et al. 2016, with Lei et al. 2016, and with the confidence and adequate justification criteria of Ehsan et al. 2019. In our experiment, we simply substitute a human user for pred and evaluate performance.

Readability and understandability. We test the user's preference for a certain style of rationale beyond their ability to predict the correct label. Our hypothesis is that humans will prefer contiguous to non-contiguous rationales. This condition aligns with coherency (Lei et al., 2016), human-likeness and understandability (Ehsan et al., 2019).

8.1 Experiments

We compare extracted rationales on two tasks, Movies and MultiRC, both of which include reference human rationales (DeYoung et al., 2020). We did not choose evidence inference for this set of experiments since the task requires expert knowledge. Recall that the rationalization task for the Movies dataset involves selecting those words or phrases associated with positive or negative sentiment. For MultiRC, the rationale must contain sufficient context to allow the user to discern whether the provided answer to the question is true, based on the information in the passage.

We extract rationales, both contiguous and non-contiguous, from 100 randomly-selected test set instances for the following methods: (1) human (reference label) rationales, (2) randomly selected rationales of length k, (3) rationales from the best Lei et al. 2016 models, and (4) rationales from the best FRESH models.

We present each extracted rationale to three annotators.[5] We ask them to perform the following tasks:

1. Classify examples as either *Positive* or *Negative* (Movies), or as *True* or *False* (MultiRC);
2. Rate their confidence on a 4-point Likert scale from *not confident* (1) to *very confident* (4);
3. Rate how easy the text is to read and understand on a 5-point Likert scale from *very difficult* (1) to *very easy* (5).

The first two tasks are designed to evaluate sufficiency, and the third readability and understandability. We provide images of the user interface in Appendix C.

We validate the user interface design with gold-label human rationales. As expected, when using these rationales Turkers are able to perform the labelling task with high accuracy, and they do so with high confidence and readability (first rows of Tables 3 and 4). On average, annotators exhibit over 84% and 89% inter-annotator agreement on Movies and MultiRC, respectively.[6]

[4] We received approval for this study from Northeastern University's Institutional Review Board (IRB).

[5] We use Amazon Mechanical Turk for the annotation task, and compensate Turkers at a rate of \$0.24 per HIT. Pay rate is calculated based on the median HIT completion time in a preliminary experiment (2 minutes) and an hourly wage of \$7.20. We require annotators to be within the U.S., but we do not explicitly test for English language proficiency.

[6] We assign the majority predicted document label and averaged Likert value for confidence and readability across the 3 annotators for each instance. We report human Accuracy as

Rationale Source	Human Acc.	Confidence (1–4)	Readability (1–5)
Human	.99	3.44 ±0.53	3.82 ±0.56
Random			
Contiguous	.84	3.18 ±0.55	3.80 ±0.57
Non-Contiguous	.65	2.09 ±0.51	2.07 ±0.69
Lei et al. 2016			
Contiguous	.88	3.39 ±0.48	4.17 ±0.59
Non-Contiguous	.84	2.97 ±0.72	2.90 ±0.88
FRESH Best			
Contiguous	.92	3.31 ±0.48	3.88 ±0.57
Non-Contiguous	.87	3.23 ±0.47	3.63 ±0.59

Table 3: Human evaluation results for Movies.

Rationale Source	Human Acc.	Confidence (1–4)	Readability (1–5)
Human	.87	3.50 ±0.47	4.16 ±0.54
Random			
Contiguous	.65	2.85 ±0.76	3.49 ±0.74
Non-Contiguous	.58	2.56 ±0.68	2.39 ±0.73
Lei et al. 2016			
Contiguous	.57	2.90 ±0.58	3.63 ±0.71
Non-Contiguous	.66	2.45 ±0.67	2.19 ±0.75
FRESH Best			
Contiguous	.69	2.78 ±0.67	3.68 ±0.6
Non-Contiguous	.65	2.60 ±0.68	2.50 ±0.83

Table 4: Human evaluation results for MultiRC.

8.2 Results

We report results in Tables 3 and 4. We observe that humans perform comparably to the trained model (Table 2) at predicting document labels given only the model-extracted rationales. Humans perform at least as well using our extracted rationales as they do with other methods. They also exhibit a strong preference for contiguous rationales, supporting our hypothesis. Lastly, we observe that confidence and readability are high. Thus while our primary goal is to provide faithful rationales, these results suggest that those provided by FRESH are also reasonably plausible. This shows that faithfulness and plausibility are not mutually exclusive, but also not necessarily correlative.

9 Conclusions

We have proposed Faithful Rationale Extraction from Saliency tHresholding (FRESH), a simple, flexible, and effective method to learn explainable neural models for NLP. Our method can be used with any feature importance metric, is very simple to implement and train, and empirically often outperforms more complex rationalized models.

FRESH performs discrete rationale selection and ensures the faithfulness of provided explanations — regardless of the complexity of the individual components — by using independent extraction and prediction modules. This allows for contextualized models such as transformers to be used, without sacrificing explainability (at least at the level of rationales). Further, we accomplish this without recourse to explicit rationale-level supervision such as REINFORCE or the reparameterization trick; this greatly simplifies training.

We showed empirically that FRESH outperforms existing models, recovering most of the performance of the original 'black-box' model. Additionally, we found FRESH rationales to be at least as plausible to human users as comparable end-to-end methods.

We acknowledge some important limitations of this work. Here we have considered explainability as an instance-specific procedure. The final explanation provided by the model is limited to the tokens provided by the extraction method. Our framework does not currently support further pruning (or expanding) this token set once the rationale has been selected.

In addition, while we do have a guarantee under our model about which part of the document was used to inform a given classification, this approach cannot readily say why this specific rationale was selected in the first place. Nor do we clearly understand how the pred uses extracted rationale to perform its classification. We view these as interesting directions for future work.

Acknowledgments

We thank colleagues at Georgia Tech and Northeastern for their feedback, including Sandeep Soni, Diyi Yang, Ian Stewart, and Jiaao Chen. We also thank the anonymous reviewers for many useful comments and suggestions.

This work was supported in part by the Army Research Office (W911NF1810328), and by the National Science Foundation (CAREER award 1750978). YP is a Bloomberg Data Science PhD Fellow.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

a measure of how well our annotators have done at predicting the correct document label from only the extracted rationale. All metrics are averaged over the 100 test documents.

learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Joost 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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Gino Brunner, Yang Liu, Damin Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, Conference Track Proceedings*.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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*, Seattle, Washington. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

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

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87. ACM.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274. ACM.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*.

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

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 01, page 282289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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 5198–5208.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Marco Tulio 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*, pages 1135–1144. ACM.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Kevin Small, Byron C Wallace, Carla E Brodley, and Thomas A Trikalinos. 2011. The constrained weight space svm: learning with ranked features. In *Proceedings of the 28th International Conference on Machine Learning*, pages 865–872. Omnipress.

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.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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 4085–4094.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using annotator rationales to improve machine learning for text categorization. In *Proceedings of the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Omar F Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 31–40.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 2016, page 795. NIH Public Access.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.

A Model Details and Hyperparameters

For each model below, we use BERT-base-uncased (for SST, AgNews), Roberta-base (for Multirc), and SciBERT (scivocab-uncased) (for Evidence Inference) embeddings (from `huggingface` library (Wolf et al., 2019) as they appear in AllenNLP library (Gardner et al., 2018)) as corresponding pretrained transformer model.

Tokenization was performed using tokenizer associated with each pretrained transformer models. Only the top two layers of each model were finetuned. For documents greater than 512 in length, we used staggered position embeddings (for example, if an example of length 1024, the position embeddings used are 1,1,2,2,3,3,...).

Lei *et al.* and Bastings *et al.* Models We use transformer model to generate token embeddings (max-pooling embeddings from wordpieces) in the generator, placing a dense classification layer on top to return a binary decision. The encoder model also uses the transformer to encode selected tokens and the start token embedding was used to perform final classification.

For the movies dataset we used a slightly different model to get around the $O(n^2)$ memory bottleneck. Specifically, we first encode 512 token subsequences with the transformer and then run these through a 128-d BiLSTM on top of transformer embeddings. Wordpiece embeddings are averaged to create token embeddings and these embeddings are then used to make token level decisions for generator model. In the encoder model, they are collapsed using additive attention module (Bahdanau et al., 2015) into a single vector prior to the final classification.

We used cross-entropy loss to train the encoder, and the optimization was performed using the Adam Optimizer with a learning rate of 2e-5. For regularization, we used 0.2 dropout after transformer embedding layer and placed an 0.001 ℓ_2 loss over all weights of our network and a grad norm of 5.0. Models were trained for 20 epochs and we kept the best parameters on the basis of macro-F1 score on dev sets.

Hyperparameter search for Lei et al. (2016) models was performed over λ_1 and λ_2 parameters, with λ_1 uniformly selected over log scale in range [1e-2, 1e-0] and λ_2 selected from [0.0, 0.5, 1.0, 2.0]. We performed the hyperparameter search 20 times and selected the best of these on the basis of F1 score on dev sets.

Bastings et al. (2019) do not require hyperparameter search since it uses a Lagrangian relaxation based optimisation for its regularizers. We use the same initial hyperparameter settings used by the authors in their codebase. We use the Hard Kumaraswamy distribution as provided by the authors here `https://github.com/bastings/interpretable_predictions`.

FRESH For all three components of the FRESH model, we used the same transformer-based models as mentioned previously to encode tokens. Classification was performed using start token embeddings. Optimisation was performed using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5. We insert a dropout layer following the BERT embedding layer for regularisation, and impose an 0.001 ℓ_2 loss over all weights of our network. We also enforce a grad norm of 5.0. The model was trained for 20 epochs, and we again kept the best models with respect to macro F1 scores on the dev sets.

For the movies dataset, we use similar modifications as discussed above.

B Hyperparameter sensitivity analysis

In Figure 7, we report the model accuracy for various hyperparameter searches on three of our datasets. Note that in many cases, the search does not converge to the desired length (it either selects the entire document or completely degenerates, selecting no tokens). We also show in Figure 6 an analysis of model performance with respect to hyperparameter search using the procedure described in (Dodge et al., 2019).

C Amazon Mechanical Turk Layouts

See Figures 8 and 9 for screenshots of the interfaces shown to annotators.

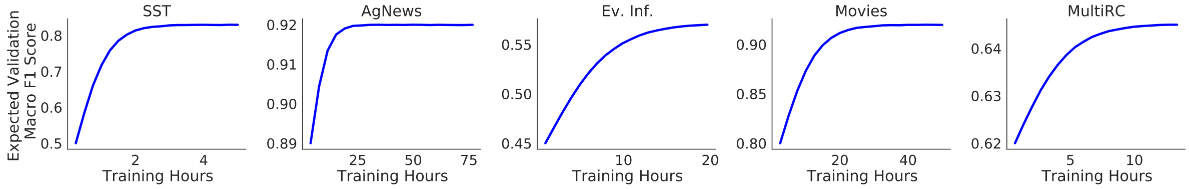


Figure 6: Plot of training duration for a single model vs Expected validation macro F1 scores as defined by (Dodge et al., 2019)

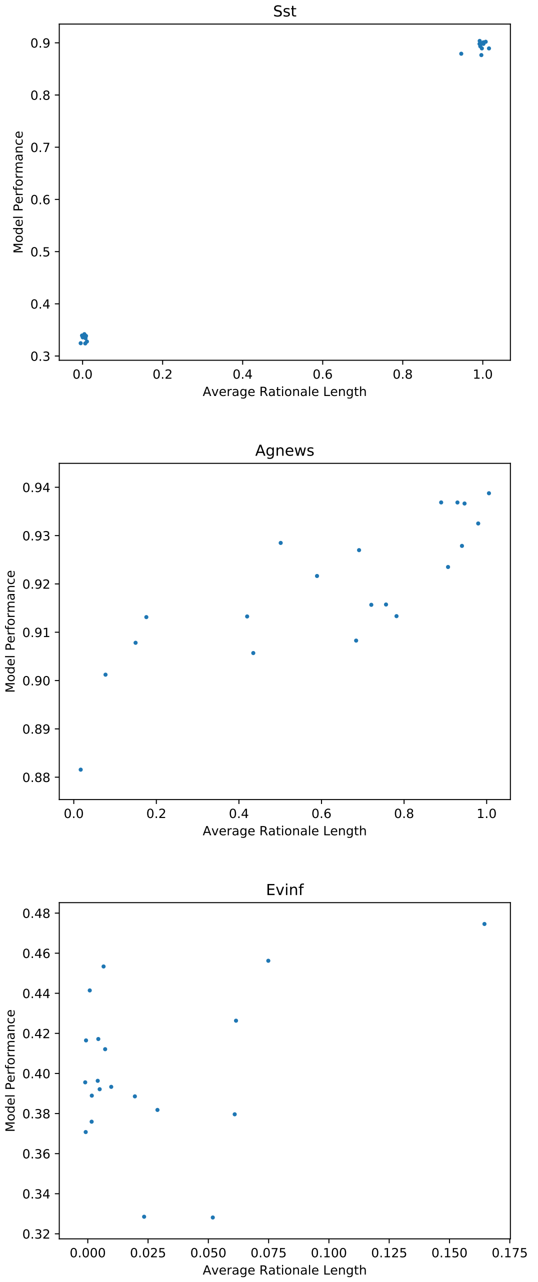


Figure 7: Plot of average rationale length vs model performance in terms of F1 score for 20 hyperparameter searches for (Lei et al., 2016)

.

D Additional Dataset Details

See Table 5.

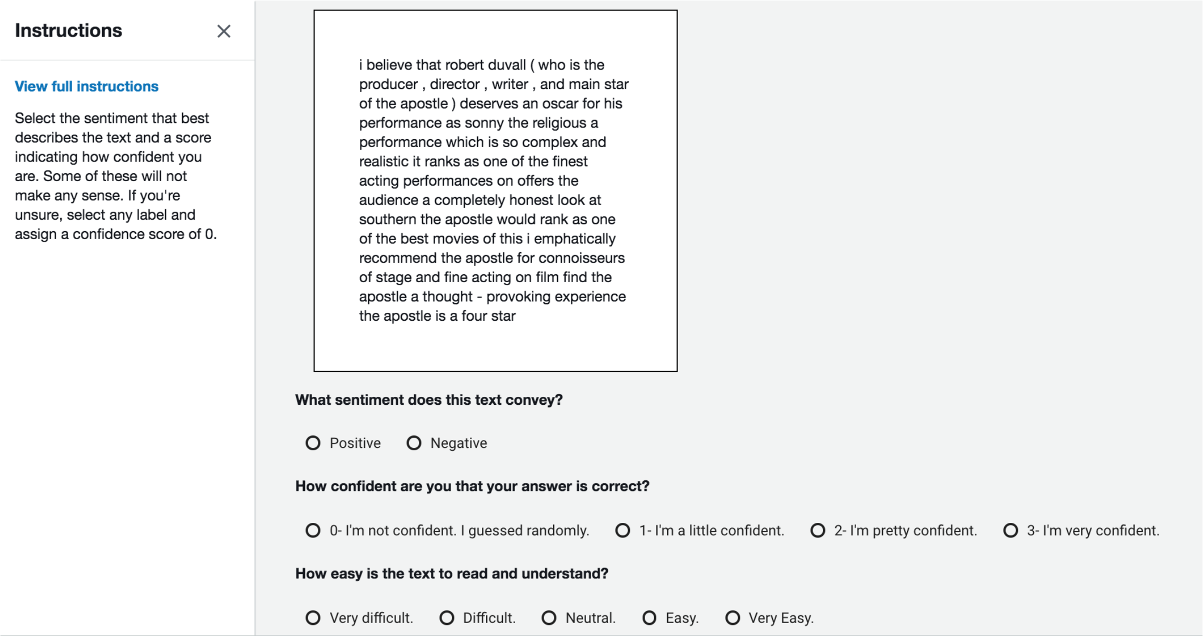
E Global Discretization Heuristics

We use following method to construct globally optimal rationales :

Global top-k. A ratio of tokens to maintain p is determined beforehand , but instead of taking the top $p \cdot |x_i|$ tokens from each instance, the training set (resp. dev, test) is created by only taking the top $p \cdot \sum_{x_i \in \mathcal{D}_{\sqcup\triangledown}} |x_i|$ tokens from the entire initial training set (resp. dev, test). To avoid the possible complete removal of certain instances, we add a further constraint where each instance first secures the top $q < p$-proportion tokens, before filling up the remainder globally.

Global Contig. To limit the rationales to a contiguous text span, we first find the maximum-mass segments for the appropriate range of lengths on each instance, then run a greedy algorithm to find per-instance lengths for overall maximal mass: starting with the minimally-long spans, of total length L_m, we perform $\mathcal{B} - L_m$ iterations of finding the next single-token addition in the entire dataset which will lead to the maximum increase in overall weight (each time ranking the marginal gains for each instance x_i of replacing the current $k(x_i)$-length span with the best $k(x_i) + 1$-length span in the instance).

In Table 6 we provide average differences between using global vs instance heuristics to extract rationales from our documents, given saliency scores. We also ran a t-test to determine if global heuristics provided results significantly different from instance-level methods, finding that they did not.



Instructions

View full instructions

Select the sentiment that best describes the text and a score indicating how confident you are. Some of these will not make any sense. If you're unsure, select any label and assign a confidence score of 0.

i believe that robert duvall (who is the producer , director , writer , and main star of the apostle) deserves an oscar for his performance as sonny the religious a performance which is so complex and realistic it ranks as one of the finest acting performances on offers the audience a completely honest look at southern the apostle would rank as one of the best movies of this i emphatically recommend the apostle for connoisseurs of stage and fine acting on film find the apostle a thought - provoking experience the apostle is a four star

What sentiment does this text convey?

○ Positive ○ Negative

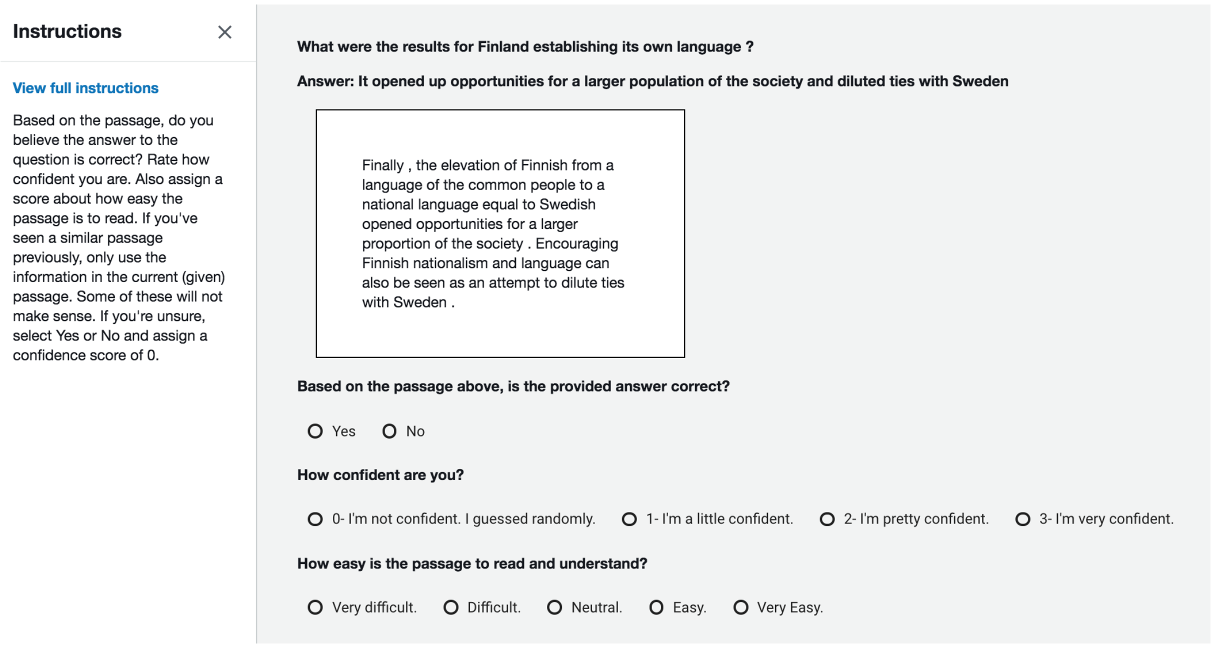
How confident are you that your answer is correct?

○ 0- I'm not confident. I guessed randomly. ○ 1- I'm a little confident. ○ 2- I'm pretty confident. ○ 3- I'm very confident.

How easy is the text to read and understand?

○ Very difficult. ○ Difficult. ○ Neutral. ○ Easy. ○ Very Easy.

Figure 8: Amazon Mechanical Turk layout for Movies tasks.



Instructions

View full instructions

Based on the passage, do you believe the answer to the question is correct? Rate how confident you are. Also assign a score about how easy the passage is to read. If you've seen a similar passage previously, only use the information in the current (given) passage. Some of these will not make sense. If you're unsure, select Yes or No and assign a confidence score of 0.

What were the results for Finland establishing its own language ?

Answer: It opened up opportunities for a larger population of the society and diluted ties with Sweden

Finally , the elevation of Finnish from a language of the common people to a national language equal to Swedish opened opportunities for a larger proportion of the society . Encouraging Finnish nationalism and language can also be seen as an attempt to dilute ties with Sweden .

Based on the passage above, is the provided answer correct?

○ Yes ○ No

How confident are you?

○ 0- I'm not confident. I guessed randomly. ○ 1- I'm a little confident. ○ 2- I'm pretty confident. ○ 3- I'm very confident.

How easy is the passage to read and understand?

○ Very difficult. ○ Difficult. ○ Neutral. ○ Easy. ○ Very Easy.

Figure 9: Amazon Mechanical Turk layout for MultiRC tasks.

	N	Doc Length	Query Length	Rationale Length	Label Distribution
Evidence Inference					
train	5,789	363 / 1010	14 / 66	0.10 / 0.54	0.39 / 0.33 / 0.28
dev	684	369 / 602	14 / 108	0.11 / 0.35	0.40 / 0.35 / 0.25
test	720	362 / 617	16 / 100	0.10 / 0.34	0.39 / 0.35 / 0.26
MultiRC					
train	24,029	305 / 618	18 / 92	0.17 / 0.73	0.56 / 0.44
dev	3,214	305 / 562	18 / 83	0.19 / 0.76	0.55 / 0.45
test	4,848	290 / 490	18 / 80	0.18 / 0.56	0.57 / 0.43
Movies					
train	1,600	773 / 2,809	7 / 7	0.09 / 0.5	0.5 / 0.5
dev	200	761 / 1,880	7 / 7	0.07 / 0.26	0.5 / 0.5
test	199	795 / 2,122	7 / 7	0.31 / 0.91	0.5 / 0.5
SST					
Train	6,920	17 / 48	-	-	0.52 / 0.48
Dev	872	17 / 44	-	-	0.51 / 0.49
Test	1,821	17 / 52	-	-	0.50 / 0.50
AgNews					
Train	102,000	31 / 173	-	-	0.25 / 0.25 / 0.25 / 0.25
Dev	18,000	31 / 168	-	-	0.25 / 0.25 / 0.25 / 0.25
Test	7,600	30 / 129	-	-	0.25 / 0.25 / 0.25 / 0.25

Table 5: Dataset statistics. For document, query, and rationale lengths we provide mean and maximum values (formulated as mean/max), where available. We do not have human rationale annotations for SST and AgNews, hence we do not report query and rationale lengths for these.

			Δ	t-statistic	p-value
dataset	saliency	rationale			
SST	Gradient	contiguous	-0.0097	-1.8483	0.1383
		Non contiguous	0.0120	1.9411	0.1242
	[CLS] Attention	contiguous	-0.0133	-3.1281	0.0352
		Non contiguous	-0.0025	-0.5036	0.6410
AgNews	Gradient	contiguous	-0.0433	-26.8053	0.0000
		Non contiguous	-0.0014	-0.7530	0.4934
	[CLS] Attention	contiguous	-0.0257	-19.4711	0.0000
		Non contiguous	-1.0000	-1.0000	-1.0000
Evidence Inference	Gradient	contiguous	-0.0126	-0.5457	0.6143
		Non contiguous	-0.0139	-0.9352	0.4026
	[CLS] Attention	contiguous	-0.0145	-1.4655	0.2166
		Non contiguous	0.0053	0.3776	0.7249
Movies	Gradient	contiguous	-0.0221	-6.4826	0.0029
		Non contiguous	-0.0020	-0.2684	0.8016
	[CLS] Attention	contiguous	-0.0232	-3.1249	0.0354
		Non contiguous	0.0040	1.6500	0.1743
MultiRC	Gradient	contiguous	-0.0041	-1.4573	0.2188
		Non contiguous	0.0066	0.8969	0.4205
	[CLS] Attention	contiguous	-0.0038	-1.1710	0.3066
		Non contiguous	0.0012	0.2832	0.7910

Table 6: Comparison of global rationales vs instance level rationale for each dataset, saliency and rationale type combination. The statistical test used was Welch's t-test (2-sided). Δ = (Average F1 score for global) - (average F1 score for instance level) heuristics.