Practical Benefits of Feature Feedback Under Distribution Shift

Anurag Katakkar*, Clay H. Yoo*, Weiqin Wang,
Zachary C. Lipton, Divyansh Kaushik
Carnegie Mellon University
akatakka,hyungony,weiqinw,zlipton,dkaushik@cmu.edu

Abstract

In attempts to develop sample-efficient and interpretable algorithms, researcher have explored myriad mechanisms for collecting and exploiting *feature feedback* (or *rationales*) auxiliary annotations provided for training (but not test) instances that highlight salient evidence. Examples include bounding boxes around objects and salient spans in text. Despite its intuitive appeal, feature feedback has not delivered significant gains in practical problems as assessed on iid holdout sets. However, recent works on counterfactually augmented data suggest an alternative benefit of supplemental annotations, beyond interpretability: lessening sensitivity to spurious patterns and consequently delivering gains in out-of-domain evaluations. We speculate that while existing methods for incorporating feature feedback have delivered negligible in-sample performance gains, they may nevertheless provide out-of-domain benefits. Our experiments addressing sentiment analysis, show that feature feedback methods perform significantly better on various natural out-of-domain datasets despite comparable in-domain evaluations. By contrast, performance on natural language inference remains comparable. Finally, we compare those tasks where feature feedback does (and does not) help.

1 Introduction

Addressing various classification tasks in natural language processing (NLP), including sentiment analysis (Zaidan et al., 2007), natural language inference (NLI) (DeYoung et al., 2020), and propaganda detection (Pruthi et al., 2020), researchers have introduced resources containing additional side information by tasking humans with marking spans in the input text (called *rationales* or *feature feedback)* that provide supporting evidence for the label. For example, spans like "underwhelming", "horrible", or "worst film since Johnny English" might indicate negative sentiment in a movie review. Conversely, spans like "exciting", "amazing", or "I never thought Vin Diesel would make me cry" might indicate positive sentiment.

These works have proposed a variety of strategies for incorporating feature feedback as additional supervision (Lei et al., 2016; Zhang et al., 2016; Lehman et al., 2019; Chen et al., 2019; Jain et al., 2020; DeYoung et al., 2020; Pruthi et al., 2020). Other researchers have studied the learning-theoretic properties of feature feedback (Poulis and Dasgupta, 2017; Dasgupta et al., 2018; Dasgupta and Sabato, 2020). We focus our study on the resources and practical methods developed for NLP.

Some have used this feedback to perturb instances for data augmentation (Zaidan et al., 2007), while others have explored multitask objectives for simultaneously classifying documents and extracting rationales (Pruthi et al., 2020). A number of papers exploit feature feedback as intermediate supervision for building extract-then-classify pipelines (Chen et al., 2019; Lehman et al., 2019; Jain et al., 2020). One common assumption is that resulting models would learn to identify and rely more on spans relevant to the target labels, which would in turn lead to more accurate predictions.

However, despite their intuitive appeal, feature feedback methods have thus far yielded underwhelming results on independent drawn and identically distributed (iid) test sets in applications involving deep nets. While Zaidan et al. (2007) found significant gains when incorporating rationales into their SVM learning scheme, benefits have been negligible in the BERT era. For example, although Pruthi et al. (2020) and Jain et al. (2020) address a different aim towards boosting interpretability—to improve extraction accuracy—their experiments show no improvement in classification accuracy by incorporating rationales.

On the other hand, Kaushik et al. (2020), introduced counterfactually augmented data (CAD) with the primary aim of showing how supplementary annotations can be incorporated to make mod-

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 346 - 355
December 8, 2022 ©2022 Association for Computational Linguistics

els less sensitive to spurious patterns, and additionally demonstrated that models trained on CAD degraded less in a collection of out-of-domain tests than their vanilla counterparts. In followup work, they showed that for both CAD and feature feedback, although corruptions to evidence spans via random word flips result in performance degradation both in- and out-of-domain, when non-evidence spans are corrupted, out-of-domain performance often improves (Kaushik et al., 2021). These findings echo earlier results in computer vision (Ross et al., 2017a; Ross and Doshi-Velez, 2018) where regularizing input gradients (so-called *local explanations*) to accord with expert attributions led to an improved out-of-domain performance.

In this paper, we conduct an empirical study of the out-of-domain benefits of incorporating feature feedback in selected domains in NLP (sentiment analysis and NLI). We seek to address two primary research questions: (i) do models that rely on feature feedback generalize better out of domain compared to *classify-only* models (i.e., models trained without feature feedback)? and (ii) do we need to solicit feature feedback for an entire dataset or can significant benefits be realized with a modest fraction of examples annotated? Our experiments on sentiment analysis (Zaidan et al., 2007) and NLI (DeYoung et al., 2020) use both linear, BERT (Devlin et al., 2019), and ELECTRA (Clark et al., 2020) models, using two feature feedback techniques (Pruthi et al., 2020; Jain et al., 2020).

We limit our experiments to sentiment analysis and NLI only, although other tasks such as hate speech and propaganda detection might appear to be natural candidates to include in our study as well. Hate Speech detection is an inherently subjective task. For example, (Waseem, 2016) documented the disagreement between labels collected from the crowd and those annotated by experts. Similarly, (Ross et al., 2017b) documented that annotating hate speech itself is a hard task leading to low inter-rater agreement within the crowd as well. Thus, even though several hate speech classification datasets exist, in our view, they are not suitable for the research questions we ask in the paper—what might be labeled as hate by one annotator may not be labeled hate in another dataset by another annotator, making it difficult to attribute the impact on performance to generalization ability or some other factors (such as noisy labeling, or choice of labeling instructions, etc.). As for propaganda detection, while a dataset with high-quality labels and feature feedback annotations exists, the lack of additional datasets restricts our ability to train and evaluate the resulting models on a battery of out-of-domain datasets.

We find that sentiment analysis models fine-tuned with feature feedback on IMDb data see no improvement in in-domain accuracy. However, out-of-domain, sentiment analysis models benefit significantly from feature feedback. For example, ELECTRA and BERT models both see gains of $\approx$ 6% on both Amazon (Ni et al., 2019) and Yelp reviews (Kaushik et al., 2021) even when feature feedback is available for just 25% of instances. However, on NLI, we find that both iid and out-of-domain performance are comparable with or without feature feedback. We further find that while for sentiment analysis, rationales constitute only $\approx$ 21% of all unique tokens in the training set, for NLI they constitute $\approx$ 80%, potentially helping to explain why feature feedback is less useful there.

2 Methods and Datasets

We focus on two techniques (classify-and-extract (Pruthi et al., 2020) and extract-then-classify (Jain et al., 2020)), two pretrained models, and one (in-domain) dataset each for sentiment analysis and NLI that contain feature feedback. For both techniques, *feature feedback* annotations provide supervision to the extractive component. The classify-and-extract model jointly predicts the (categorical) label and performs sequence tagging predict rationales. The classification head and a linear chain CRF (Lafferty et al., 2001) share an encoder, initialized with pretrained weights.

The extract-then-classify method (Jain et al., 2020) first trains a classifier (*support*) on complete examples to predict the label, using its outputs to extract continuous feature importance scores. These scores are then binarized using a second classifier (*extractor*) which is trained on the feature importance scores from *support* and makes token-level binary predictions to identify rationale tokens in the input. A binary cross-entropy term in the objective of the extractor is used to maximise agreement of the extracted tokens with human rationales. Finally, a third classifier (*predictor*) is trained to predict the target (sentiment or entailment) label based only on these extracted tokens.

For both approaches, we experiment with two pretrained models (BERT and ELECTRA). We

Test set	Classify-only	Pruthi et al.	Jain et al.
BERT			
In-domain	$85.9_{0.7}$	$\mathbf{89.9_{2.3}}$	$\mathbf{90.4_{0.3}}$
CRD	$89.3_{0.7}$	$\mathbf{91.6_{0.7}}$	$87.5_{0.8}$
SST2	$77.6_{4.1}$	$79.3_{3.6}$	$75.6_{1.2}$
Amazon	$78.1_{4.9}$	$83.5_{3.1}$	$\mathbf{92.3_{1.2}}$
Semeval	$70.6_{5.7}$	$73.2_{2.6}$	$68.6_{2.2}$
Yelp	$86.8_{1.7}$	$85.7_{1.6}$	$\mathbf{91.6_{0.1}}$
ELECTRA			
In-domain	$93.2_{0.3}$	$91.8_{1.4}$	$93.1_{0.3}$
CRD	$91.6_{0.4}$	$\mathbf{93.7_{0.9}}$	$91.5_{0.7}$
SST2	$73.2_{1.3}$	$74.0_{1.2}$	$\mathbf{77.2_{1.4}}$
Amazon	$72.8_{2.0}$	$75.5_{2.1}$	$\mathbf{84.2_{1.6}}$
Semeval	$67.5_{4.5}$	$72.5_{1.8}$	$66.7_{3.0}$
Yelp	$79.0_{3.6}$	$\mathbf{84.6_{1.8}}$	$\mathbf{94.7_{0.2}}$

Table 1: Mean and standard deviation (in subscript) of accuracy scores of classify-only models, and models proposed by Pruthi et al. (2020) and Jain et al. (2020), fined-tuned for sentiment analysis. Significant results ($p < 0.05$) compared to the classify-only models are highlighted in bold.

limit the maximum sequence length to 512 tokens and train all models for 10 epochs using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2e-5$ and a batch size of 8 and early stopping based on mean of classification and extraction F1 scores on the validation set. We replicate all experiments on 5 seeds and report mean performance along with standard deviation.

To see whether results are consistent across architectures, we also use a linear SVM (Zaidan et al., 2007) with a modified objective function on top of the ordinary soft-margin SVM, i.e.,

$$\frac{1}{2}||w||^2 + C(\sum_i \delta_i) + C_{\text{contrast}}(\sum_{i,j} \xi_{ij})$$

subject to the constraints $\vec{w} \cdot \vec{x}_{ij} \cdot y_i \geq 1 - \xi_{ij}\ \forall i, j$ where $\vec{x}_{ij} := \frac{\vec{x}_i - \vec{v}_{ij}}{\mu}$ are *psuedoexamples*, created by subtracting *contrast-examples* ($\vec{v}_{ij}$), input sentence void of randomly chosen rationales, from the original input ($\vec{x}_i$). We use term-frequency embeddings with unigrams appearing in at least 10 reviews and set $C = C_{contrast} = \mu = 1$. For each training example, we generate 5 psuedoexamples.

Datasets For sentiment analysis, we use an IMDb movie reviews dataset (Zaidan et al., 2007). Reviews in this dataset are labeled as having either *positive* or *negative* sentiment. Zaidan et al. (2007) also tasked annotators to mark spans in each review that were indicative of the overall sentiment. We use these spans as feature feedback. Overall, the dataset has 1800 reviews in the training set (with feature feedback) and 200 in test (without feature feedback). Since the test set does not include ground truth labels for evidence extraction, we construct a test set out of the 1800 examples in the original training set. This leaves 1200 reviews for a new training set, 300 for validation, and 300 for test. For NLI, we use a subsample of the E-SNLI dataset (DeYoung et al., 2020) used in Kaushik et al. (2021). In this dataset, there are 6318 premise-hypothesis pairs, equally divided across *entailment* and *contradiction* categories.

Test set	Classify-only	Pruthi et al.	Jain et al.
BERT			
In-domain	$88.7_{2.0}$	$89.8_{0.8}$	$77.7_{0.1}$
RP	$62.9_{3.9}$	$66.6_{0.6}$	$57.9_{0.1}$
RH	$76.9_{3.5}$	$80.5_{1.9}$	$70.7_{0.2}$
MNLI-M	$69.7_{2.6}$	$68.1_{1.9}$	$69.8_{0.1}$
MNLI-MM	$71.5_{2.7}$	$69.2_{2.3}$	$66.2_{0.1}$
ELECTRA			
In-domain	$\mathbf{96.0_{0.2}}$	$95.0_{0.3}$	$85.4_{0.04}$
RP	$80.8_{1.0}$	$78.0_{0.6}$	$72.2_{0.1}$
RH	$88.9_{1.0}$	$88.7_{0.9}$	$79.7_{0.1}$
MNLI-M	$86.5_{0.9}$	$81.9_{2.1}$	$77.1_{0.1}$
MNLI-MM	$86.6_{0.8}$	$82.1_{2.0}$	$75.7_{0.1}$

Table 2: Mean and standard deviation (in subscript) of F1 scores of models fine-tuned for NLI with an increasing number of examples with feature feedback. Significant results ($p < 0.05$) compared to the classify-only models are highlighted in bold.

We evaluate on CRD (Kaushik et al., 2020), SST-2 (Socher et al., 2013), Amazon reviews (Ni et al., 2019), Tweets (Rosenthal et al., 2017) and Yelp reviews (Kaushik et al., 2021) for sentiment analysis, and Revised Premise (RP), Revised Hypothesis (RH) (Kaushik et al., 2020), MNLI matched (MNLI-M) and mismatched (MNLI-MM) (Williams et al., 2018) for NLI.

3 Experiments

We first fine-tune BERT and ELECTRA on the annotated IMDb dataset (Zaidan et al., 2007) following both classify-and-extract and extract-then-classify approaches. We evaluate resulting models on both iid test set as well as various naturally occurring out-of-domain datasets for sentiment analysis and compare resulting performance with classify-only models (Table 1). We find that both approaches lead to significant gains (when tested with t-test with $p < 0.05$) in out-of-domain performance compared to the classify-only method. For

Evaluation set	No rationales	Fraction of Training Data with Rationales 25%	50%	75%	100%
BERT					
In-domain	$85.9_{0.7}$	$\mathbf{87.7_{1.1}}$	$88.1_{2.4}$	$\mathbf{90.2_{1.5}}$	$\mathbf{89.9_{2.3}}$
CRD	$89.3_{0.7}$	$\mathbf{91.7_{0.6}}$	$\mathbf{92.3_{0.9}}$	$\mathbf{92.3_{0.3}}$	$\mathbf{91.6_{0.7}}$
SST2	$77.6_{4.1}$	$81.2_{0.6}$	$81.3_{0.7}$	$81.8_{0.6}$	$79.3_{3.6}$
Amazon	$78.1_{4.9}$	$\mathbf{85.3_{1.2}}$	$\mathbf{84.6_{1.7}}$	$\mathbf{84.0_{0.5}}$	$83.5_{3.1}$
Semeval	$70.6_{5.7}$	$\mathbf{77.8_{1.0}}$	$75.5_{0.8}$	$74.9_{0.8}$	$73.2_{2.6}$
Yelp	$86.8_{1.7}$	$86.9_{1.1}$	$85.8_{1.5}$	$85.4_{0.7}$	$85.7_{1.6}$
ELECTRA					
In-domain	$93.2_{0.3}$	$92.4_{0.9}$	$92.8_{1.2}$	$93.7_{1.9}$	$91.8_{1.4}$
CRD	$91.6_{0.4}$	$92.1_{0.8}$	$\mathbf{93.0_{0.6}}$	$\mathbf{93.1_{0.3}}$	$\mathbf{93.7_{0.9}}$
SST2	$73.2_{1.3}$	$73.1_{1.8}$	$72.3_{1.6}$	$72.3_{1.1}$	$74.0_{1.2}$
Amazon	$72.8_{2.0}$	$\mathbf{79.0_{1.8}}$	$\mathbf{75.7_{1.2}}$	$\mathbf{76.6_{1.8}}$	$75.5_{2.1}$
Semeval	$67.5_{4.5}$	$70.5_{1.5}$	$66.2_{1.5}$	$67.1_{2.2}$	$72.5_{1.8}$
Yelp	$79.0_{3.6}$	$\mathbf{84.5_{1.1}}$	$\mathbf{84.2_{1.7}}$	$\mathbf{84.3_{1.2}}$	$\mathbf{84.6_{1.8}}$

Table 3: Mean and standard deviation (in subscript) of accuracy scores of models fine-tuned for sentiment analysis using the method proposed by Pruthi et al. (2020) with different base models (BERT and ELECTRA) and increasing proportion of examples with feature feedback. Results highlighted in bold are significant difference with $p < 0.05$.

Evaluation set	No rationales	Fraction of Training Data with Rationales 25%	50%	75%	100%
BERT					
In-domain	$88.7_{2.0}$	$89.6_{0.4}$	$89.9_{0.4}$	$89.7_{0.4}$	$89.8_{0.8}$
RP	$62.9_{3.9}$	$67.6_{2.0}$	$67.4_{1.2}$	$68.6_{0.6}$	$66.6_{0.6}$
RH	$76.9_{3.5}$	$80.4_{1.1}$	$81.7_{1.6}$	$81.4_{0.7}$	$80.5_{1.9}$
MNLI-M	$69.7_{2.6}$	$67.6_{3.4}$	$68.1_{4.6}$	$68.8_{2.0}$	$68.1_{1.9}$
MNLI-MM	$71.5_{2.7}$	$68.8_{4.5}$	$69.2_{5.9}$	$69.8_{2.7}$	$69.2_{2.3}$
ELECTRA					
In-domain	$\mathbf{96.0_{0.2}}$	$95.1_{0.3}$	$95.0_{0.3}$	$95.0_{0.3}$	$95.0_{0.3}$
RP	$80.8_{1.0}$	$78.2_{1.3}$	$79.2_{1.1}$	$77.2_{1.3}$	$78.0_{0.6}$
RH	$88.9_{1.0}$	$88.0_{1.2}$	$88.4_{0.3}$	$87.9_{0.4}$	$88.7_{0.9}$
MNLI-M	$86.5_{0.9}$	$82.0_{2.8}$	$82.4_{1.6}$	$82.3_{0.9}$	$81.9_{2.1}$
MNLI-MM	$86.6_{0.8}$	$82.6_{2.8}$	$83.5_{1.4}$	$82.6_{0.8}$	$82.1_{2.0}$

Table 4: Mean and standard deviation (in subscript) of F-1 scores of models fine-tuned for NLI using the method proposed by Pruthi et al. (2020) with different base models (BERT and ELECTRA) and increasing proportion of examples with feature feedback. Results highlighted in bold are significant difference with $p < 0.05$.

instance, ELECTRA fine-tuned using the extract-then-classify framework leads to $\approx 15.7\%$ gain in accuracy when evaluated on Yelp. For NLI, however, training with rationales doesn't lead to any visible performance gain (Table 2).

As Pruthi et al. (2020) demonstrate better performance on evidence extraction for sentiment analysis compared to Jain et al. (2020), we use their method for additional analysis. For both sentiment analysis and NLI, we fine-tune models with varying proportion of samples with rationales and report iid and out-of-domain performance (Tables 3 and 4). Training with no feature feedback recovers the classify-only baseline.

On sentiment analysis, we find feature feedback to improve BERT's iid performance but find ELEC-

Test set	Classify-only	Zaidan et al.
In-domain	$75.2_{3.5}$	$79.1_{3.4}$
CRD	$48.3_{2.0}$	$\mathbf{58.2_{2.4}}$
SST-2	$49.7_{0.3}$	$\mathbf{65.6_{1.5}}$
Amazon	$50.9_{0.3}$	$\mathbf{68.7_{3.1}}$
Semeval	$49.8_{0.1}$	$\mathbf{58.0_{1.5}}$
Yelp	$55.7_{2.8}$	$\mathbf{74.8_{2.7}}$

Table 5: Mean and standard deviation (in subscript) of accuracy scores of classify-only SVM model versus SVM trained with feature feedback for sentiment analysis using Zaidan et al. (2007)'s method. Significant results ($p < 0.05$) compared to the classify-only models are highlighted in bold.

Task	Unigram	Bigram
Sentiment Analysis	21.37	11.20
NLI	79.54	35.49

Table 6: Percentage of unigram and bigram vocabularies that are marked as feature feedback at least once.

	Entailment	Contradiction
D_{all}	0.25	0.16
$D_{\text{rationale}}$	0.30	0.09

Table 7: Mean Jaccard index of premise-hypothesis word overlap (D_{all}) and rationale overlap ($D_{\text{rationale}}$) in the training set.

Dataset	% Overlap	Label Agreement
	Unigram	
CRD	60.3	51.3
SST2	64.6	66.5
Amazon	45.6	47.6
Semeval	30.9	60.3
Yelp	78.3	65.1
	Bigram	
CRD	28.2	51.9
SST2	28.5	64.5
Amazon	19.6	49.9
Semeval	10.2	58.5
Yelp	46.8	65.3

Table 8: Rationale vocabulary overlap and label agreement between in-sample and OOD datasets.

TRA's performance comparable with and without feature feedback. Feature feedback leads to an increase in performance out-of-domain on both BERT and ELECTRA. For instance, with feature feedback, ELECTRA's classification accuracy increases from 91.6% to 93.7% on CRD and 79% to 84.6% on Yelp. Similar trends are also observed when we fine-tune BERT with feature feedback. Interestingly, when evaluated on the SemEval dataset (Tweets), we observe that BERT fine-tuned with feature feedback on all training examples achieves comparable performance to fine-tuning without feature feedback. However, fine-tuning with feature feedback on just 25% of training examples leads to a significant improvement in classification accuracy. We speculate that this might be a result of implicit hyperparameter tuning when combining prediction and extraction losses, and a more extensive hyperparameter search could provide comparable (if not better) gains with 100% data. Similarly, SVM trained with feature feedback (Zaidan et al., 2007) consistently outperformed SVM trained without feature feedback, when evaluated out-of-domain despite obtaining similar accuracy in-domain (Table 5 and Appendix Table 11). For instance, SVM trained on just label information achieved $75.2\% \pm 3.5\%$ accuracy on the in-domain test set, which was comparable to the accuracy of $79.1\% \pm 3.4\%$ achieved by SVM trained with feature feedback. But the classifier trained with feature feedback led to $\approx 19\%$ and $\approx 18\%$ improvement in classification accuracy on Yelp reviews and Amazon reviews, respectively, compared to the classifier trained without feature feedback.

For NLI, it appears that feature feedback provides no added benefit compared to a classify-only BERT model, whereas, ELECTRA's iid performance decreases with feature feedback. Furthermore, models fine-tuned with feature feedback generally perform no better than classify-only models when trained with varying proportions of rationales (Table 4) while classify-only models perform significantly better than the models trained with rationales when trained with varying dataset size. (Appendix Table 2). These results are in line with observations in prior work on counterfactually augmented data (Huang et al., 2020).

4 Discussion and Analysis

To further study the different trends on sentiment analysis versus NLI, we analyze feature feedback in both datasets. We find that 21.37% of tokens in the vocabulary of Zaidan et al. (2007) are marked as rationales in at least one movie review. Interestingly, this fraction is 79.54% for NLI (Table 6). While for movie reviews, certain words or phrases might generally denote positive or negative sentiment (e.g., "amazing movie"), for NLI tasks, it is not clear that any individual phrase should suggest entailment or contradiction generally. A word or a phrase might be marked as indicating entailment in one NLI example but as a contradiction in another. This may explain why training with rationales lead to no improvement in the NLI task.

We further construct vocabulary of unigrams and bigrams from phrases marked as feature feedback in examples from the sentiment analysis training set ($V_{\text{rationale}}$). We compute the fraction of unigrams (and bigrams) that occur in this vocabulary and also occur in each out-of-domain dataset. We find that a large fraction of unigrams from $V_{\text{rationale}}$ also exist in CRD ($\approx 60\%$), SST2 ($\approx 64\%$), and Yelp ($\approx 78\%$) data. (movie and restaurant reviews). However, this overlap is much smaller for SemEval ($\approx 30\%$) and Amazon ($\approx 45\%$) , which consist of tweets and product reviews, respectively. For these overlapping unigrams, we observe a relatively large percentage (50–65%) preserve their associated majority training set label in the out-of-domain datasets. Similar trends hold for bigrams, though fewer $V_{\text{rationale}}$ bigrams are present out-of-domain (Table 8). A model that pays more attention to these spans might perform better out of domain.

For each pair in the NLI training set, we compute Jaccard similarity between the premise and hypothesis sentence (Table 7). We compute the mean of these example-level similarities over the entire dataset, finding that it is common for examples in our training set to have overlap between premise and hypothesis sentences, regardless of the label. However, when we compute mean Jaccard similarity between premise and hypothesis rationales, we find higher overlap for entailment examples versus contradiction. Thus, models trained with feature feedback might learn to identify word overlap as predictive of entailment even when the true label is contradiction. While this may not improve an NLI model's performance, it could be useful in tasks like Question Answering, where answers often lie in sentences that have high word overlap with the question (Lamm et al., 2020; Majumder et al., 2021). Interestingly, our results on NLI are in conflict with recent findings where models trained with rationales showed significant improvement over classify-only models in both iid and out-of-domain (MNLI-M and MNLI-MM) settings (Stacey et al., 2021). This could be due to the different modeling strategy employed in their work, as they use rationales to guide the training of the classifier's attention module. Investigating this difference is left for future work.

5 Conclusion

In this paper, we investigate the practical benefits of using feature feedback in two well-known tasks in NLP: sentiment analysis and natural language inference. Using two techniques that were primarily introduced for boosting interpretability as the basis of our experiments, we find they also have an unexpected advantage in boosting model robustness. Our experiments and analyses offer insight into how these interpretability methods may encourage generalization in out of domain settings.

To answer our first research question, we show that models trained with feature feedback can lead to performance improvement in the sentiment analysis task but not in NLI. To answer our second question, we find that as little as 25% of the dataset can achieve the best performance in the out-of-domain setting in sentiment analysis, whereas no clear trends are visible in NLI. Our analysis reveals that a smaller percentage of vocabulary is selected as rationales in sentiment analysis compared to NLI, indicating rationale tokens in the sentiment analysis task contain more distinctive information than NLI. Rationale tokens are more likely to exist among entailment samples than contradiction, which may lead the model to correlate the existence of rationales with entailment.

References

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, and Sivan Sabato. 2018. Learning from discriminative feature feedback. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Sanjoy Dasgupta and Sivan Sabato. 2020. Robust learning from discriminative feature feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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 *Association for Computational Linguistics (ACL)*.

William Huang, Haokun Liu, and Samuel Bowman. 2020. Counterfactually-augmented snli training data does not yield better generalization than unaugmented data. In *First Workshop on Insights from Negative Results in NLP*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. In *Association for Computational Linguistics (ACL)*.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering. *arXiv preprint arXiv:2009.06354*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Sagnik Majumder, Chinmoy Samant, and Greg Durrett. 2021. Model agnostic answer reranking system for adversarial question answering. In *European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL SRW)*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Stefanos Poulis and Sanjoy Dasgupta. 2017. Learning with feature feedback: from theory to practice. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Weakly-and semi-supervised evidence extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *International Workshop on Semantic Evaluation (SemEval)*.

Andrew Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI Conference on Artificial Intelligence*.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017a. Right for the right reasons: training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017b. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *ArXiv*, abs/1701.08118.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2021. Natural language inference with a human touch: Using human explanations to guide model attention. *arXiv preprint arXiv:2104.08142*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *NLP+CSS@EMNLP*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies: North American chapter of the association for computational linguistics (NAACL-HLT)*.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Empirical Methods in Natural Language Processing (EMNLP)*.

A Appendix

Task	Examples
Sentiment Analysis (Positive)	. . . characters are portrayed with such saddening realism that you can't help but love them , as pathetic as they really are . although levy stands out , guest , willard , o'hara , and posey are all wonderful and definitely should be commended for their performances ! if there was an oscar for an ensemble performance , this is the group that should sweep it . . .
Sentiment Analysis (Negative)	. . . then , as it's been threatening all along , the film explodes into violence . and just when you think it's finally over , schumacher tags on a ridiculous self-righteous finale that drags the whole unpleasant experience down even further . trust me . there are better ways to waste two hours of your life . . .
NLI (Entailment)	**P:** a white dog drinks water on a mountainside. **H:** there is a dog drinking water right now.
NLI (Contradiction)	**P:** a dog leaping off a boat **H:** dogs drinking water from pond

Table 9: Examples of documents (and true label) with feature feedback (highlighted in yellow).

Task	Examples
Sentiment Analysis (Positive, Correct)	everyone should adapt a tom robbins book for screen . while the movie is fine and the performances are good , the dialogue , which works well reading it , is beautiful when spoken .
Sentiment Analysis (Positive, Wrong)	... very uncaptivating yet one gets the feeling that their is some serious exploitation going on here ...
Sentiment Analysis (Negative, Correct)	... using quicken is a frustrating experience each time i fire it up ...
Sentiment Analysis (Negative, Wrong)	... with many cringe-worthy 'surprises', which happen around 10 minutes after you see exactly what's going to happen ...
NLI (Entailment, Correct)	**P:** a woman cook in an apron is smiling at the camera with two other cooks in the background . **H:** a woman looking at the camera .
NLI (Entailment, Wrong)	**P:** a woman in a brown dress looking at papers in front of a class . **H:** a woman looking at papers in front of a class is not wearing a blue dress .
NLI (Contradiction, Correct)	**P:** the woman in the white dress looks very uncomfortable in the busy surroundings **H:** the dress is black .
NLI (Contradiction, Wrong)	**P:** a man , wearing a cap , is pushing a cart , on which large display boards are kept , on a road . **H:** the person is pulling large display boards on a cart .

Table 10: Examples (from out-of-domain evaluation sets; with true label and model prediction) of explanations highlighted by feature feedback models (highlighted in yellow).

Evaluation Set	Dataset size 300	600	900	1200
In-domain	$77.0_{3.9}$/$77.6_{2.2}$	$78.5_{3.2}$/$82.3_{2.0}$	$80.5_{1.7}$/$\mathbf{84.9_{1.6}}$	$75.2_{3.5}$/$79.1_{3.4}$
CRD	$48.0_{2.9}$/$\mathbf{56.4_{1.3}}$	$48.3_{2.5}$/$\mathbf{58.0_{2.7}}$	$48.4_{2.3}$/$\mathbf{58.7_{1.8}}$	$48.3_{2.0}$/$\mathbf{58.2_{2.4}}$
SST-2	$52.2_{1.6}$/$\mathbf{62.9_{1.0}}$	$50.9_{3.0}$/$\mathbf{64.0_{0.9}}$	$51.3_{3.1}$/$\mathbf{64.9_{0.9}}$	$49.7_{0.3}$/$\mathbf{65.6_{1.5}}$
Amazon	$51.8_{1.5}$/$\mathbf{65.9_{1.9}}$	$52.4_{2.0}$/$\mathbf{66.5_{1.2}}$	$52.0_{2.9}$/$\mathbf{69.9_{0.4}}$	$50.9_{0.3}$/$\mathbf{68.7_{3.1}}$
Semeval	$50.3_{1.4}$/$\mathbf{56.7_{1.1}}$	$50.3_{1.2}$/$\mathbf{56.4_{0.8}}$	$50.1_{0.5}$/$\mathbf{58.8_{1.3}}$	$49.8_{0.1}$/$\mathbf{58.0_{1.5}}$
Yelp	$60.2_{4.0}$/$\mathbf{72.0_{2.4}}$	$57.3_{7.1}$/$\mathbf{74.5_{1.5}}$	$61.2_{4.6}$/$\mathbf{74.8_{2.5}}$	$55.7_{2.8}$/$\mathbf{74.8_{2.7}}$

Table 11: Mean and standard deviation (in subscript) of accuracy scores of classify-only SVM models (left) presented alongside accuracy scores of models trained with feature feedback (right), with increasing number of training-samples for sentiment analysis using the method proposed by Zaidan et al. (2007). Results highlighted in bold show statistically significant difference with $p < 0.05$.

Evaluation Set	Dataset size 1500	3000	4500	6318
BERT				
In-domain	$85.9_{6.0}$/$84.5_{2.0}$	$87.9_{0.4}$/$87.7_{1.0}$	$89.1_{0.4}$/$89.2_{0.2}$	$88.7_{2.0}$/$89.8_{0.8}$
RP	$61.8_{0.9}$/$62.8_{1.8}$	$63.3_{1.6}$/$64.2_{1.8}$	$63.7_{1.8}$/$\mathbf{66.8_{1.4}}$	$62.9_{3.9}$/$66.4_{1.7}$
RH	$74.5_{1.6}$/$71.8_{3.4}$	$77.0_{1.4}$/$77.3_{2.1}$	$78.3_{1.1}$/$80.4_{1.8}$	$76.9_{3.5}$/$80.5_{1.9}$
MNLI-M	$63.7_{3.1}$/$60.8_{3.2}$	$69.2_{1.8}$/$66.3_{2.2}$	$70.2_{0.9}$/$67.5_{3.1}$	$69.7_{2.6}$/$68.1_{1.9}$
MNLI-MM	$64.8_{4.3}$/$61.8_{4.3}$	$71.3_{2.3}$/$67.5_{2.8}$	$72.1_{1.2}$/$68.9_{4.2}$	$73.1_{1.9}$/$71.4_{1.1}$
ELECTRA				
In-domain	$\mathbf{94.6_{0.2}}$/$92.7_{0.5}$	$\mathbf{95.1_{0.4}}$/$94.2_{0.3}$	$\mathbf{95.7_{0.2}}$/$94.4_{0.2}$	$\mathbf{96.0_{0.2}}$/$95.1_{0.3}$
RP	$78.4_{1.2}$/$75.2_{2.5}$	$78.5_{1.8}$/$77.2_{0.9}$	$\mathbf{81.2_{0.6}}$/$76.2_{1.2}$	$\mathbf{80.8_{1.0}}$/$78.0_{0.6}$
RH	$\mathbf{87.7_{0.7}}$/$85.2_{1.4}$	$88.1_{1.3}$/$87.3_{0.6}$	$\mathbf{89.4_{0.6}}$/$87.1_{1.0}$	$88.9_{1.0}$/$88.7_{0.9}$
MNLI-M	$\mathbf{82.8_{2.2}}$/$77.0_{1.8}$	$\mathbf{85.4_{1.8}}$/$78.9_{1.7}$	$\mathbf{86.0_{1.6}}$/$80.4_{2.1}$	$\mathbf{86.5_{0.9}}$/$81.9_{2.1}$
MNLI-MM	$\mathbf{83.6_{2.5}}$/$77.9_{2.1}$	$\mathbf{86.2_{2.1}}$/$79.9_{1.9}$	$\mathbf{86.1_{1.8}}$/$80.8_{2.2}$	$\mathbf{86.6_{0.8}}$/$82.1_{2.0}$

Table 12: Mean and standard deviation (in subscript) of F-1 scores of classify-only models/models trained with feature feedback, with increasing number of training-samples for NLI using the method proposed by Pruthi et al. (2020). Results highlighted in bold are statistically significant difference with $p < 0.05$.