Investigation and Mitigation of Dataset Artifacts Using Model Ablations and Adversarial Challenge Sets

Shawyan Mozaffar[1], Ernest Petrilli[1]
[1]Natural Language Processing, The University of Texas at Austin

Abstract

This paper investigates and addresses dataset artifacts in Natural Language Inference (NLI) using model ablations and adversarial challenge sets. It explores how state-of-the-art NLP models, while proficient in tasks like evaluation on the Stanford Natural Language Inference (SNLI) dataset, can inadvertently learn and rely on unintended dataset artifacts, leading to biased or skewed results. The study employs the TextAttack framework and an ELECTRA-small model to identify and quantify these artifacts. It further proposes an approach of data augmentation and model fine-tuning for improving model robustness, particularly towards neutral-labeled examples. This work contributes to enhancing model accuracy and generalizability in NLI tasks.

1 Introduction

1.1 Background

Natural Language Inference (NLI) is a task in the realm of Natural Language Processing (NLP) that focuses on categorizing a semantic relationship, or lack thereof, between a natural language sequence premise *p* and a natural language sequence hypothesis *h*. If the hypothesis is supported by the premise, entailment is inferred, if the hypothesis opposes the premise, contradiction is inferred, and neutrality is inferred when the relationship between the premise and hypothesis remains undetermined (MacCartney, 2009; MacCartney and Manning, 2008). Rather than focusing on sequences of formal reasoning and deduction, NLI emphasizes informal reasoning, lexical semantic knowledge, and variability of linguistic expression (MacCartney, 2009).

Entailment	*h* is definitely true given *p*
Neutral	*h* might or might not be true given *p*
Contradiction	*h* is definitely not true given *p*

Table 1: Criteria used for generating three hypotheses given a premise sentence for crowd sourcing SNLI hypotheses.

1.2 SNLI Dataset for NLI

The lack of diverse, large datasets of labeled natural language inference data was addressed by Bowman et al. (2015), who used large-scale crowd-sourcing annotations to create the Stanford Natural Language Inference (SNLI) dataset. The SNLI consists of 570K (550K train pairs, 10K dev pairs, and 10K test pairs) human-labeled premise-hypothesis pairs which are categorized as entailment, neutral, or contradiction. The premises in the corpus were sourced from the Flickr30k corpus, a set of 160k crowdsourced image captions which reflect a range of human experiences and contexts and provide a foundation for constructing complex NLI scenarios (Bowman et al., 2015; Young et al., 2014). A team of Amazon Mechanical Turk workers were then tasked with generating three hypotheses for a premise according to the criteria shown in Table 1 (Bowman et al., 2015; Gururangan et al., 2018):

Attack Recipe	Base - Attack Success Rate %	Fine Tuned - Attack Success Rate %	Transformation	Capability Tested
bae	77.89	**73.53**	BERT masked token prediction	Sensitivity to context-aware word replacements
bert-attack	100	100	BERT masked token prediction with subword expansion	Sensitivity to context-aware word replacements
checklist	3.16	3.19	Contracts, extends, substitutes names entities	Invariance to name entity changes and sentence modifications
clare	100	100	RoBERTa masked prediction for token swap, insert, merge	Sensitivity to token-level perturbations
deepwordbug	100	**96.81**	Character insertion, deletion, swap	Sensitivity to character-level perturbations and typos
fast-alzantot	77.89	**74.81**	Counter-fitted word embedding swap	Invariance to semantic word changes
iga	98.95	**98.94**	Counter-fitted word embedding swap	Invariance to algorithm-based word substitutions
input-reduction	100	100	Word deletion	Dependency on specific words for prediction
kuleshov	100	100	Counter-fitted word embedding swap	Invariance to vector-based word changes
pruthi	74.74	**70.21**	Character-level swap, deletion, insertion	Invariance to common typographical errors
pso	44.21	47.87	HowNet word swap	Sensitivity to semantic changes using HowNet
pwws	98.95	100	WordNet-based synonym swap	Handling of word importance and synonym substitutions
textbugger	100	**97.87**	Character-level swap, deletion, insertion	Robustness to charater-level attacks in a black-box setting
textfooler	97.89	**97.82**	Counter-fitted word embedding swap	Sensitivity to semantically similar word replacements

Table 2: Types of transformations and model capabilities tested for each TextAttack recipe. Attack success rate shown per recipe for the base ELECTRA-small model and the fine-tuned ELECTRA-small model.

1.3 Dataset Artifacts

State-of-the-art NLP models can achieve high performance on NLI datasets such as the SNLI (Bowman et al., 2015), however, these models may also inadvertently learn certain unintended patterns, biases, or cues that may be introduced during the data annotation process, known as *dataset artifacts* or *annotation artifacts*. These artifacts can skew the model's learning process, resulting in an overestimation of semantic language understanding capabilities and a reduction in generalizability. For example, annotators might consistently use certain words in the hypotheses they construct for entailment or contradiction, which can lead models to predict the inference label based on these superficial cues rather than actual semantic reasoning (Gururangan et al., 2018).

Poliak et al. (2018) showed that models trained solely on hypotheses, without the premises, can perform unexpectedly well, indicating that the hypotheses contain enough information for the model to predict the correct label; this is evidence that dataset artifacts exist in the hypotheses.

1.4 Adversarial Challenge Sets

Several methods for assessing and mitigating dataset artifacts have become prominent in NLI research. One method, adversarial challenge sets, involves designing new inputs that resemble the training data yet are subtly modified to change their meaning. These challenge sets are meticulously crafted with the intent to expose the vulnerabilities of models and reveal their over-reliance on dataset-specific cues and patterns, which indicate a lack of natural language understanding (Jia & Liang, 2017; Wallace et al., 2019). The examples within these sets often include counterfactuals or minimal pairs – sentences that are almost identical but differ in one critical aspect that changes the model's prediction (such as a word or common phrase) (Glockner et al., 2018). Studies by Mccoy et al. (2019) and Bartolo et al. (2020) show that adversarial challenge sets can reveal when models are leveraging syntactic heuristics rather than using linguistic comprehension to make a genuine inference. Liu et al. (2019) found that models trained on examples from adversarial challenge sets perform better on standard NLI benchmarks and show improved generalization to novel contexts.

1.5 TextAttack Framework

The TextAttack framework offers a structured and efficient method for performing adversarial attacks on NLP models in order to better understand the potential presence of dataset artifacts; it allows for the identification and quantification of the degree to which models are influenced by these artifacts and enables the investigation of whether a model's predictions are based on robust language understanding or if they are merely capitalizing on the unintended cues and patterns encoded within the training data (Morris et al., 2020). TextAttack is a python library which hosts an arsenal of 'recipes,' which are various methods for transforming input text under specific constraints. Table 2 includes the TextAttack recipes used, and their respective transformations applied, and capabilities tested. By introducing subtle variations to the input data that maintain the semantic meaning while superficially differing from the data seen during

Dataset	Accuracy	Input for Model Training
SNLI	89.68	Premise-Hypothesis Pair
SNLI	69.91	Hypothesis-only

Table 3: Accuracy of the ELECTRA-small model trained on full SNLI dataset compared to accuracy of ablated ELECTRA-small model trained and evaluated on hypothesis-only examples.

Dataset	Accuracy	Data Augmentation
SNLI	89.68	No
SNLI	**90.17**	TextAttack Recipes

Table 4: Accuracy of the ELECTRA-small model evaluated on the SNLI test set compared to accuracy of fine-tuned ELECTRA-small model evaluated on SNLI test set.

training, if a model's performance deteriorates significantly on these transformed inputs, it signals that the model may be overfitting to specific lexical patterns rather than comprehending the underlying semantics of the text (Wallace et al., 2019).

2 Investigation of Dataset Artifacts

2.1 Detection of Artifacts

For the investigation and mitigation of dataset artifacts, we used the Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECRTRA-small) model as our pre-trained model (Clark et al., 2020). ELECTRA has the same architecture as BERT, however, during training the model replaces some input tokens with plausible alternatives sampled from a small generator network rather than replacing input tokens with a mask. It is also computationally efficient compared to larger models. To establish a baseline, we trained the ELECTRA model on the entirety of the SNLI train set, and then evaluated the trained model on the entirety of the SNLI test set. We also replicated the results of Poliak et al. (2018) by using model ablations to remove the premises in the SNLI train set and trained the ELECTRA model again on the hypothesis-only inputs. Upon evaluation using hypothesis-only examples in the SNLI test set, we found that the model performed remarkably well using just the hypotheses, further reinforcing the finding that certain patterns and cues exist in the hypotheses which provide adequate information for the model to make correct predictions. Table 3 shows the accuracies from the model trained on the full SNLI and the ablated model trained on hypothesis-only examples.

2.2 Exploratory Analysis of SNLI

Upon further examination of the base model trained on the full SNLI dataset, we observed that 42.32% of all the incorrect predictions were when the gold label was neutral. In our improvements, we set out to mitigate dataset artifacts and improve the model by training it to appropriately predict the neutral label more often rather than predicting the entailment or contradiction labels more often. Exploratory analysis of the SNLI dataset revealed that the neutral label was not overrepresented, as less than 1/3rd (32.87%) of all the gold labels were neutral, implying that the model struggled more with accurately predicting the neutral gold label (Table 7).

On reviewing examples of the TextAttack recipes, it became evident that we needed to interpret whether the successful attacks resulted in an appropriate update to the previous prediction. Because of our previous findings about the tendency of the model to overinterpret, we focused on examples where the model shifted from predicting neutral to either entailment or contradiction.

After manually inspecting whether the new predictions seemed appropriate, we observed that only the Clare recipe, which perturbed tokens, consistently resulted in appropriate new predictions by the model. For example: the premise "*Children [smiling] and waving at camera*" was perturbed to "*Children [crying] and waving at camera.*" The hypothesis remained: "*They are smiling at their parents.*" This result changed the prediction from neutral to contradiction. To remain consistent with the previous gold label, one can assume that "*they*" refers to "*children*," but we still cannot infer whether their parents are holding the camera.

Gold Label	Predicted Label	Count Corrected	Count Uncorrected	Accuracy	Total Sum	Representation of Gold Label
Entailment	Neutral	50	44	53.19%	141	26.86%
	Contradiction	23	24	48.94%		
Neutral	Entailment	57	44	56.44%	223	42.48%
	Contradiction	65	57	51.18%		
Contradiction	Entailment	21	16	56.76%	161	30.67%
	Neutral	71	53	57.26%		

Table 5: Counts of corrected examples and new incorrect examples with total accuracy of changed predictions for combinations of gold and predicted label from the fine-tuned ELECTRA-small model trained on the augmented dataset. Sum of the counts and distribution of the gold labels in augmented dataset.

However, we can infer that the model correctly considers smiling and crying to be antithetical.

Other recipes, such as BAE, which made context aware word replacements based on cosine similarity, caused the model to make inappropriate new predictions. The shared premise: "*High fashion ladies wait outside a tram beside a crowd of people in the city*" with a hypothesis that was perturbed from "*The women enjoy having a [good] fashion sense*" to "*The women enjoy having a [solid] fashion sense*" caused the model to unexpectedly switch from a neutral to contradiction prediction. The model does not seem to understand that we cannot infer whether the women enjoy their fashion sense, regardless of whether they have "good" or "solid" fashion sense.

3 Mitigation of Dataset Artifacts

3.1 TextAttack Recipes and Perturbations

Using the TextAttack framework, we used the same 100 examples from the SNLI train set across 14 different recipes to attack the input text. With the intention of causing the model to make an erroneous prediction, each recipe applied a different type of perturbation to the input text to form the adversarial challenge examples. If a perturbation to the data caused the model to change its prediction, the attack was classified as successful, and if a perturbation caused the model to maintain its original prediction, the attack was considered a failure. The higher the attack success rate, the more sensitive the model is to that specific type of perturbation, and the lower the attack success rate, the more robust the model is to that type of perturbation. Table 2 shows the attack success rate for the different TextAttack recipes using the base model trained on the SNLI.

3.2 Improving Accuracy with Augmentation

Upon the analysis of the base model and the TextAttack recipes, we propose a data augmentation strategy for the mitigation of dataset artifacts and the improvement of robustness towards neutral labeled examples. We implemented word-level augmentation using the perturbations applied in the TextAttack recipes to generate an adversarial challenge set for model fine-tuning. We hand-annotated 100 perturbed neutral examples – each of which had been successfully attacked using TextAttack recipes – with their original gold labels and added them to the SNLI train set. After fine-tuning the model on increasing intervals of adversarial challenge examples, culminating in 106 new hand-annotated examples, we observed an improvement in accuracy on the SNLI test set compared to the base model (Figure 3). Table 4 shows the accuracies observed upon evaluation of the base model and the fine-tuned model trained on the augmented data. With a 0.49% improvement in accuracy, we investigated the behavior of the model on a granular level.

4 Results

4.1 Evaluation of Models

We aggregated both the correct and incorrect predictions upon evaluation of both the base ELECTRA-small model and the fine-tuned model, as well as the disjoint sets of incorrect predictions of both models. Table 5 shows the six different categories of incorrect predictions. If the data augmentation only improved the

Gold Label	Predicted Label	Count	Net Improved	Percent Improved	Total Sum	Accuracy
Entailment	Neutral	222	2	0.89%	3329	91.35%
	Contradiction	66	0	0%		
	Entailment	3041				
Neutral	Entailment	199	16	7.44%	3235	87.36%
	Contradiction	210	5	2.33%		
	Neutral	2826				
Contradiction	Entailment	66	5	7.04%	3278	91.76%
	Neutral	204	21	9.33%		
	Contradiction	3008				

Table 6: Count of incorrect predictions for combinations of gold and predicted labels with net improvement and percent improved for each combination from the fine-tuned ELECTRA-small model trained on the augmented dataset. Sum of the counts and accuracy per gold label in augmented dataset.

robustness of the model towards neutrally labeled examples, then by comparing the incorrect predictions of the base model with those of the fine-tuned model, we would expect to see the neutral gold label achieve the highest accuracy improvement among all the gold labels.

4.2 Analysis of Errors

While we did observe a greater number of corrections on the two error categories with a neutral gold label, we also observed a greater number of examples which were uncorrected (correct on base model, incorrect on fine-tuned model). The net count of incorrect predictions, however, was decreased and the overall accuracy on the SNLI test set improved. Additionally, we observed that the accuracy for examples with contradiction gold labels improved the most compared to examples with neutral or entailment gold labels (Table 6, Table 7).

5 Discussion

The model became more accurate when it predicted neutral than any other labels (determined by aggregating the incorrect examples per predicted label), however, it became more accurate for examples with a gold label of contradiction compared to other gold labels (determined by aggregating the incorrect examples per gold label). To further investigate this, we examined the distributions of both predicted and gold labels among the adversarial challenge examples we added to the SNLI during fine-tuning. Out of the 106 (originally neutrally labeled) added examples, 74 were predicted contradiction and 32 were predicted entailment. We added 87 examples with a neutral label, 19 examples with a contradiction label, and no examples with an entailment label during the data augmentation process. Since our objective was to make the model more robust to neutral labeled examples and less sensitive to predicting entailment or contradiction, we compared the fine-tuned model's behavior on neutral examples that were predicted contradiction against contradiction examples that were predicted neutral. We did the same comparison for neutral-entailment and entailment-contradiction pairs and found that the model became more robust to predicting neutral when not provided enough evidence to suggest entailment. More surprisingly, the model became more sensitive to predicting the contradiction label. To better understand why that might have happened, we also observed that the model became more accurate in appropriately predicting contradiction for previous examples that it had predicted entailment, instead of the other way around. Out of the 49 corrected examples, only two of them were from the error categories that neither had a neutral gold label nor a neutral predicted label. Even after correcting for the distribution of the dataset, these error categories were still underrepresented in the corrected examples, namely, the model did not improve a single prediction with an entailment gold label and contradiction predicted label.

We next looked for specific corrected examples from the fine-tuned model where the model

Gold Label	Predicted Label	Count	Total Sum	Accuracy	Representation of Gold Label	Distribution of Errors
Entailment	Neutral	224	3329	91.29%	33.82%	28.54%
	Contradiction	66				
	Entailment	3039				
Neutral	Entailment	215	3235	86.71%	32.87%	42.32%
	Contradiction	215				
	Neutral	2805				
Contradiction	Entailment	71	3278	90.97%	33.31%	29.13%
	Neutral	225				
	Contradiction	2982				

Table 7: Count of incorrect predictions for combinations of gold and predicted labels from the base ELECTRA-small model trained on the SNLI. Sum of the counts and accuracy per gold label, as well as the distribution of each gold label in the SNLI train set.

appropriately switched from predicting neutral to contradiction and sought to understand them better considering the capabilities that were tested with the TextAttack recipes. Clare had been the only recipe that we overwhelmingly annotated with contradiction gold labels, and this recipe specifically tested for sensitivity to token-level perturbations. Recall the example from the Clare recipe that was mentioned earlier about the premise changing from "*smiling*" to "*crying*" children, which would now contradict the hypothesis about smiling children, even though we still don't know if it's their parents who are holding the camera. This shows that a single inconsistency between the premise and hypothesis should cause a contradiction. This is a lower burden of proof than entailment where all aspects of the hypothesis must be supported by the premise and could explain why such distinctive features of a contradiction are easier for a model to learn, perhaps with much less data. The following example that had been corrected after our dataset augmentation demonstrates this point:

Premise	A group of young men are raking lives in front of a home
Hypothesis	The men are watering roses
Gold Label	Contradiction
Previously Predicted Label	Neutral

Figure 1: Example from fine-tuned model shows a corrected example.

The model might now understand that raking and watering are two different activities that one cannot perform at the same time, just as smiling and crying are generally considered incompatible behaviors that were reinforced with our fine-tuned model. Because Clare made these perturbations on impact words, it effectively made the model more sensitive to these slight variations that caused the entire hypothesis to become a contradiction.

On the other hand, we also wanted to better understand how we achieved the intended robustness against entailment. For an example such as the following, we can appreciate how the model unlearned the spurious correlation of the same word "*concentration*" occurring in both the premise and hypothesis, while not having any support for "*crossing levels*" that was introduced in the hypothesis.

Premise	This boy is in deep concentration playing a game of ping-pong
Hypothesis	The boy crossed several levels by playing with sincerity and full concentration
Gold Label	Neutral
Previously Predicted Label	Entailment

Figure 2: Example from fine-tuned model shows improvement on certain cues.

To better illustrate our point about the relationship of robustness and sensitivity to entailment and contradiction, imagine we had the same unsupported claim about crossing levels, but with lack of focus, now the hypothesis would become a contradiction despite still containing neutral elements. In other words, a hypothesis can have any number of neutral claims and become a contradiction considering a single contradiction amongst those claims, whereas entailment would require support for all the claims. This could explain why the model

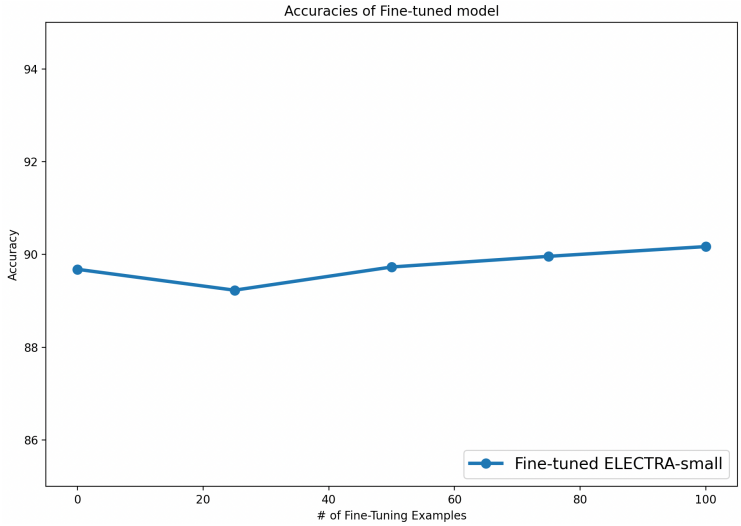

Figure 3: Fine-tuning the ELECTRA-small model on an increasing number of adversarial challenge examples. Overall model accuracy upon evaluation on SNLI test set displayed at each step.

became more accurate given a neutral prediction than given other predictions, while also performing better on contradiction with respect to the other gold labels, because it may have learned that higher standard of implication for entailment compared to contradiction.

Another way to analyze these findings is that the net impact of our improved predictions was greater than 7% of the original incorrect predictions for the following error categories: neutral examples that were predicted entailment, contradiction examples that were predicted entailment, and contradiction examples that were predicted neutral, and less than 2.3% of the original incorrect predictions for the other error categories (Table 6). It was unexpected that the model would improve as much on contradiction examples that were predicted entailment. The predictions that regressed to a new inaccurate prediction were also least for this category, suggesting that the model may have slightly overfit to some of our augmented, neutral gold label examples. Supporting the conviction that the model is still struggling the most with the ambiguity surrounding neutral examples, our greatest error category among the new incorrect predictions involved neutral gold labels with contradiction as the predicted label, demonstrating that the model learned how to remain robust to more neutral gold labels in the face of entailment predicted labels than contradiction predicted labels. This also may have resulted from an over-representation of contradiction for the predicted label in the successful attacks.

References

MacCartney, B. 2009. Natural Language Inference. Ph.D. Thesis, Stanford University.

MacCartney, B., & Manning, C. D. 2008. Modeling semantic containment and exclusion in natural language inference. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 105–112.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67–78.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. 2018. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 107–112.

Jia, R., & Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Glockner, M., Shwartz, V., & Goldberg, Y. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL).

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., & Stenetorp, P. 2020. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. Transactions of the Association for Computational Linguistics.

McCoy, T. R., Pavlick, E., & Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Proceedings of the Association for Computational Linguistics (ACL).

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. arXiv preprint arXiv:2005.05909.

Wallace, E., et al. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. EMNLP.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR).