Analyzing and Mitigating Dataset Artifacts in NLI

S.R.
MSCS, UT Austin

Abstract

Natural Language Inference (NLI) is the classification of relationship between a pair of sentences into a label. Pre-trained models for this purpose can pick up spurious artifacts from the datasets that they are trained on. Such models perform classification well on the data similar to the trained dataset but perform very weakly on slightly different dataset. In this paper we analyze the dataset artifacts that a model may learn by training this model with just one part of the pair along with the label to predict a label. We further investigate the specific errors and behavior from the model and discuss the general class of mistakes that the model makes.

We then try to improve the NLI model by exposing the model to challenging datasets called adversarial data and evaluate its performance with two different implementation approaches.

1 Introduction

Natural Language Inference is a fundamental task in Natural Language Processing, aimed at determining the relationship between a premise and a hypothesis[1](Kalouli et al.,2019). It is a vital aspect of NLP and has a variety of applications [2] (Dagan et al., 2013). Pre-trained models can usually perform well on the datasets that they are trained on. But they achieve surprisingly low performance on examples that are constructed very similar to those in training data. In the paper [3](Gardner et al., 2020) contrast examples are constructed by modifying examples in a small way. checklist examples are utilized in [4](Ribeiro et al.2020) and adversarial examples are utilized in [5](Jia and Liang, 2017).

These papers show that a model can achieve good performance on a given dataset by learning spurious correlations that are called dataset artifacts. Models that learn such dataset artifacts mostly fail on generalized examples from the real-world that do not contain these artifacts.

In this paper we attempt to understand the presence of such artifacts in a dataset by analyzing its effect on performance with adversarial data and improve a model to reduce the effects of dataset artifacts.

2 Research Background

2.a. For understanding relationship between pairs of sentences, we looked the datasets that bear similarity to NLI tasks. Question-Answering is also similar to NLI where it is trying to predict a label based of whether the given answer is in line with the question posed. Based on this we took four datasets into consideration: MultiNLI [6](Williams et

al.2018), SQuAD [7](Rajpurkar et al.2016), HotpotQA [8](Yang et al., 2018) and the Stanford NLI [9](Bowman et al., 2015).

2.b. To conduct analysis about the presence of dataset artifacts in the dataset chosen, a model had to be chosen for to be trained and evaluated for classification. We decided to use pretrained models for this purpose for which the following models were taken into consideration : BERT [10] (Devlin et al.), RoBERTa [11] (Liu et al., 2019), XLNet [12](Yang et al., 2019), ALBERT [13](Lan et at., 2019) and ELECTRA [14](Clark et al., 2020).

The following papers were also taken into consideration to understand these models and evaluate options for the suitability of the datasets with the models : [15](Raffel et al., 2023), [16](Radford et al.,2018), [17](Brown et al., 2020), [18] (Kwiatkowski et al., 2019).

2.c. To conduct analysis, the options that were considered are:

2.c.(i) Changing data sets
[19](Wallace et al., 2019), [20](Bartolo et al., 2020), [21](Glockner et al., 2018), [22] (McCoy et al., 2019)

2.c.(ii) Conducting statistical tests
[23] (Dror et al., 2018)

2.c.(iii) Using model ablations
[24](Chen and Durrett, 2019), [25](Kaushik and Lipton, 2018) [26](Poliak et al.,2018)

2.d. Research on methods to improve the model involved looking into the following options:

2.c.(i) Training on adversarial data to present a diverse range of sentences to the model where the model will fail due to presence of artifacts and learn new relationships.
[27](Zhou and Bansal, 2020), [28](Morris et al., 2020), [29] (Dua et al.2021)

2.c.(ii) Focusing on difficult subsets of the dataset and attempt to improve the accuracy of the model based on these subsets.
[30] (Swayamdipta et al.2020)

3 Data source and Materials

The **Stanford Natural Language Inference (SLNI)** dataset is used for the NLI task. The SLNI contains around 570,000 English sentence pairs containing premise and hypothesis relations. Each of these pairs is assigned a classification based on the relation between them as an Entailment, Neutral, or Contradiction. The dataset is generated from Flickr30k dataset image captions taken as premises and crowd-sourced human-elicited workers created respective hypotheses for the captions. This dataset was created by [1] Bowman et al. (2015).

The dataset is split into three sections, around 550,000 premise-hypothesis pairs are used in training the model. Test and validation sets are made of 10,000 premise-hypothesis pairs. [31](Rudinger et al., 2017) has shown that SNLI dataset contains stereotypical biases based on gender, race, and ethnic stereotypes.

The dataset is available at https://nlp.stanford.edu/projects/snli/

4 Research and Methods

4.a. Model used

For analyzing the dataset, we use the Google Electra small discriminator model. ELECTRA is a method for self-supervised language representation learning, that is used to pre-train transformer networks. This model can be easily integrated with transformers library provided by Google's Hugging Face and it simplifies the usage in NLI task. It is a popular choice due to its efficient training and good results.

The model was published by [14] (Clark et al., 2020) and is available at https://huggingface.co/google/electra-small-discriminator

4.b. Analysis of Dataset Artifacts

On training the SNLI dataset's training data with the Google Small Electra model, it achieves an evaluation loss of 0.384 and an evaluation accuracy of 0.891 on the evaluation set.

We analyze the dataset artifacts in the dataset using model ablation with a hypothesis-only NLI, which disregards the premise and tries to predict the label from the hypothesis-premise relation pairs. For this the model used in [26](Poliak et al.,2018) is used on the dataset. The code for this implementation is available at https://github.com/azpoliak/hypothesis-only-NLI

The model is based on [32] Conneau et al. (2017)'s InferSent method which uses a BiLSTM encoder and constructs a representation for a sentence by max-pooling over its hidden states which uses the sentence to classify the label. This model achieves a 0.691 evaluation accuracy on the SNLI dataset. Preprocessing is done using an NLTK tokenizer (Loper et al. 2002) and is mapped into 300-dimensional Glove vectors [33] Pennington et al. 2014).

The accuracy of the hypothesis-only model suggests the presence of considerable artifacts in the dataset. The evaluation dataset of SNLI was considered to conduct this analysis, the following aspects were considered for analysis:

4.b.(i) Length Distribution Analysis:

We examined the distribution of hypothesis lengths for each label. The average length of the hypothesis shows that it can act as an artifact:

	Entailment	Neutral	Contradict
Average Length	6.81	8.34	7.39

Table 2 - Average length of hypothesis corresponding to a label

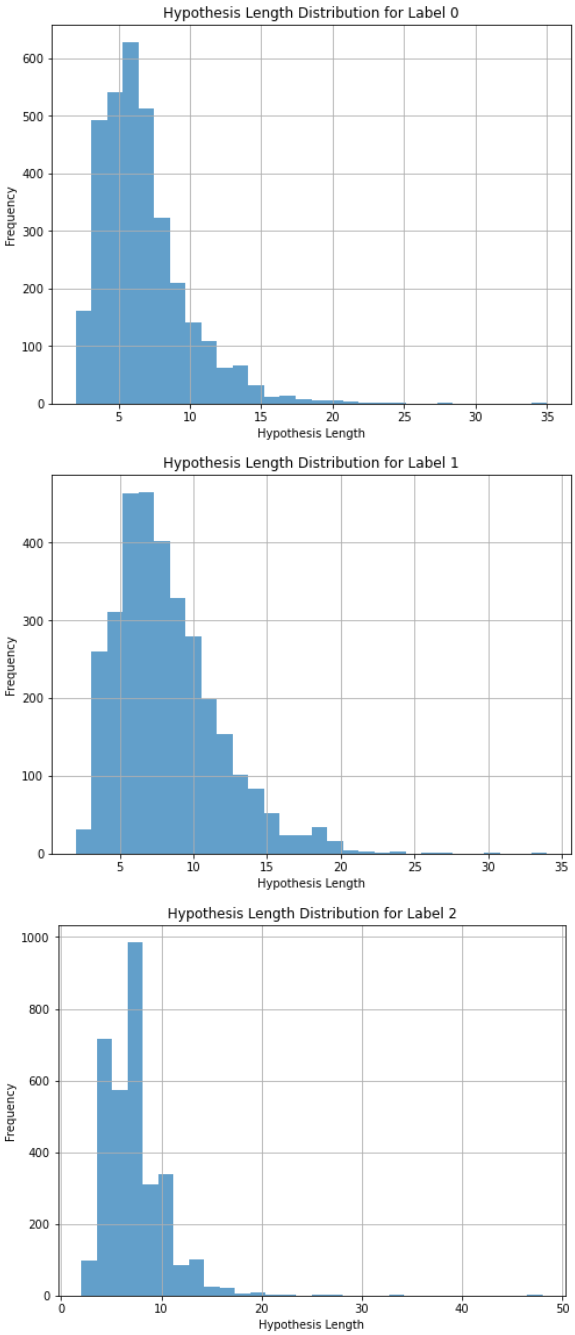


Figure 1 - Distribution of hypothesis length

Label	Mean	Median	Std dev	Min	Max	Count
Entailment (0)	6.81	6	2.93	2	35	3329
Neutral (1)	8.34	8	3.36	2	34	3235
Contradict (2)	7.39	7	2.89	2	48	3278

Table 1 - Statistics on hypothesis length distribution

We perform a Pearson correlation between the length of the hypothesis and the label and get

a correlation of 0.0774 for the validation set, which is not very significant.

We then evaluate the model's performance on subsets of data categorized by length on the e valuation dataset. We observe that the average accuracy for short hypotheses (length betwee n 0 and 10) is 0.89, accuracy for medium hyp otheses (length between 11 and 20) is 0.89, an d accuracy for long hypotheses (length betwe en 21 and inf) is 0.73.

We then perform error analysis by length to see if certain lengths are more prone to errors. This can indicate a length-based bias. We can see that errors are comparatively higher for length 16 and 21, which signals towards a length related bias.

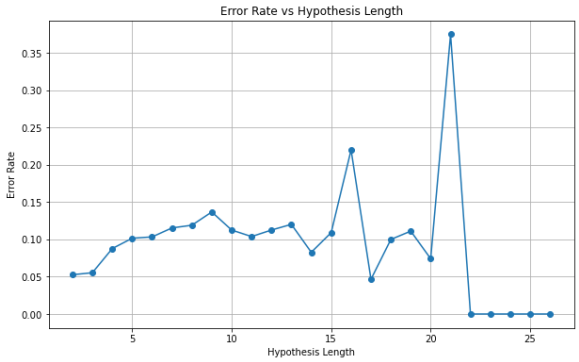


Figure 3 – Error rate versus hypothesis length

To understand the effect of lexical diversity we calculate the Type-Token-Ratio(TTR) for different lengths of the hypothesis. The ratio of unique tokens to total number of tokens in a sentence is taken for this. The decreasing TTR with increasing hypothesis length can act as an artifact which can result in the model to learn to associate length with certain labels. A consistently low or high TTR for a label can become an artifact.

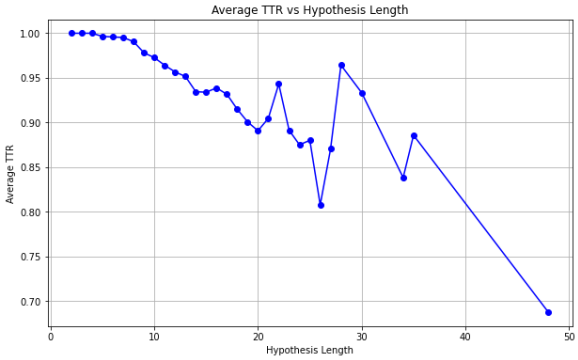


Figure 2 – Trend of Average Type Token Ratio versus hypothesis length

Shannon entropy is another measure of randomness (unpredictability) in a distribution of words in a sentence. High entropy implies the unpredictable presence of words in the sentence. The model can learn to associate a particular entropy associated with a certain length to produce an artifact. The entropy appears to increase with increase in length of the hypothesis which is due to the longer sentences having more words. This indicates that the sentences include diverse linguistic expression reducing the risk of artifacts.

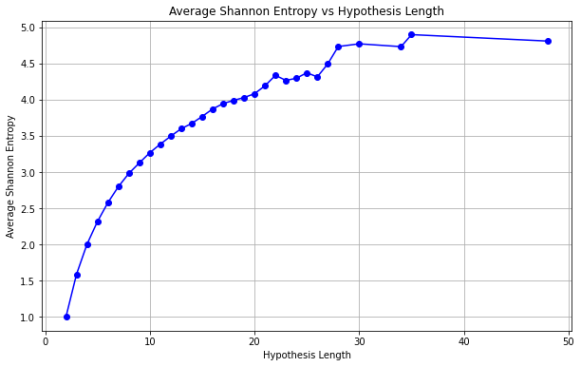


Figure 4 - Average Shannon entropy versus hypothesis length

4.b.(ii) Effect of words present in the hypothesis:

We look at the most frequent words that are associated with a particular label. The following (word, frequency) is the 20 most common for each of the labels below:

Label 0, Entailment:

```
('man', 677), ('people', 354), ('wom
an', 341), ('outside', 236), ('Two',
225), ('person', 188), ('playing', 1
88), ('girl', 159), ('wearing', 147)
```

, ('men', 143), ('People', 139), ('boy', 138), ('dog', 120), ('group', 113), ('sitting', 107), ('walking', 103), ('standing', 97), ('child', 95), ('water', 91), ('near', 88)

Label 1, Neutral:

('man', 674), ('woman', 337), ('people', 222), ('Two', 172), ('girl', 162), ('playing', 150), ('men', 144), ('boy', 136), ('dog', 127), ('person', 106), ('group', 101), ('wearing', 99), ('young', 91), ('women', 88), ('child', 78), ('outside', 78), ('walking', 76), ('two', 73), ('People', 70), ('game', 65)

Label 2, Contradict:

('man', 702), ('woman', 387), ('Two', 219), ('people', 209), ('men', 185), ('playing', 174), ('sitting', 173), ('girl', 172), ('dog', 147), ('boy', 147), ('wearing', 100), ('women', 97), ('standing', 96), ('sleeping', 95), ('group', 95), ('young', 93), ('person', 88), ('running', 80), ('eating', 80), ('two', 79)

We can see that some of the words are present for all labels, so we further look at the 20 most common distinct words that are unique to each label, and their frequency, which can act as an artifact.

Label 0 unique words:

('opposing', 5), ('physical', 4), ('liquid', 4), ('photographing', 3), ('motion', 3), ('autumn', 3), ('motorcyle', 3), ('accessories', 3), ('speaks', 3), ('hills', 3), ('stumbles', 2), ('deere', 2), ('worked', 2), ('laboratory', 2), ('involved', 2), ('worship', 2), ('physically', 2), ('movement', 2), ('podium', 2), ('blows', 2)

Label 1 unique words:

('vacation', 13), ('celebrating', 10), ('happily', 10), ('halloween', 9), ('tour', 7), ('last', 7), ('meet', 6), ('discuss', 5), ('upcoming', 5), ('likes', 5), ('members', 5), ('important', 5), ('tournament', 5), ('paid', 4), ('related', 4), ('boating', 4), ('plans', 4), ('beating', 4), ('spot', 4), ('soon', 4)

Label 2 unique words:

('nobody', 53), ('sleep', 12), ('nothing', 11), ('anything', 8), ('napping', 7), ('mars', 6), ('sofa', 6), ('spaceship', 5), ('porch', 5), ('ignoring', 5), ('cow', 5), ('ordering', 5), ('nude', 5), ('throw', 4), ('sound', 4), ('sheep', 4), ('highway', 4), ('library', 4), ('sing', 4), ('bat', 4)

In the evaluation dataset, the number of unique words that come under each label in the evaluation set with 6838 unique words of hypothesis are as follows:

	Unique words	Percentage
Label 0	489	7.15
Label 1	1264	18.48
Label 2	956	13.98
Total	2709	39.61

Table 3 - Presence of words unique to each label in hypothesis

We can see that about 40% of the total tokens can be associated uniquely with a label and this can act as an artifact in the hypothesis-only inference. We then analyze how many correct predictions are made in the presence of these unique words:

Unique word label	Number of hypotheses with Correct predictions	Percentage
Label 0	365	3.70
Label 1	1056	10.72
Label 2	899	9.13
Total	2320	23.57

Table 4 - Presence of unique words in hypothesis for labels with correct predictions

We see those hypotheses corresponding to unique words associated with a unique label form about 23.57% of the total hypothesis present in the evaluation set which indicates the considerable contribution of the presence of artifacts that the model could have learnt.

4.b.(iii) Analysis of Errors:

We analyze the effect of negation present in the hypothesis by looking at the cross tab to see how often errors occur in the presence or absence of negation in the evaluation dataset.

Contains Negation/ Is Error	False	True
False	8716	1062
True	56	5

Table 5 - Error Matrix for the presence of negation and prediction on SNLI's validation set

When we look at the label-wise negation errors in the evaluation dataset, we see a very minor fraction of the errors:

	Negation Related Error	Other Error
Label 0	1	304
Label 1	2	442
Label 2	2	316

Table 6 - Count of prediction errors connected to presence of negation in SNLI's validation set

The confusion matrix below visualizes the distribution of errors in the evaluation set.

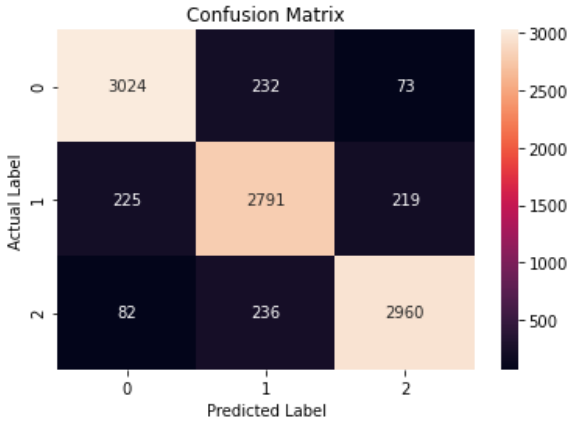
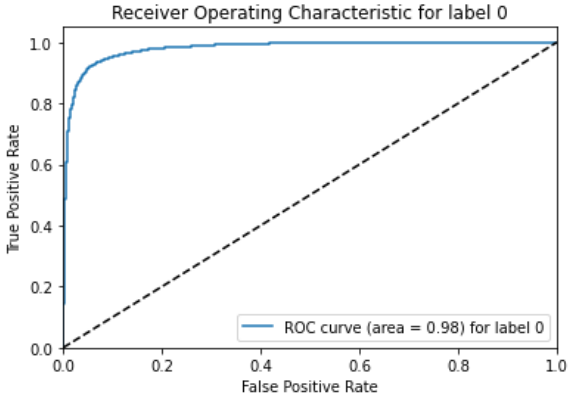


Figure 5 - Label-wise distribution of errors from SNLI validation data prediction

The ROC curves show the performance of the classification of the model at different classification thresholds. The curves show that the model performs well on the evaluation set.



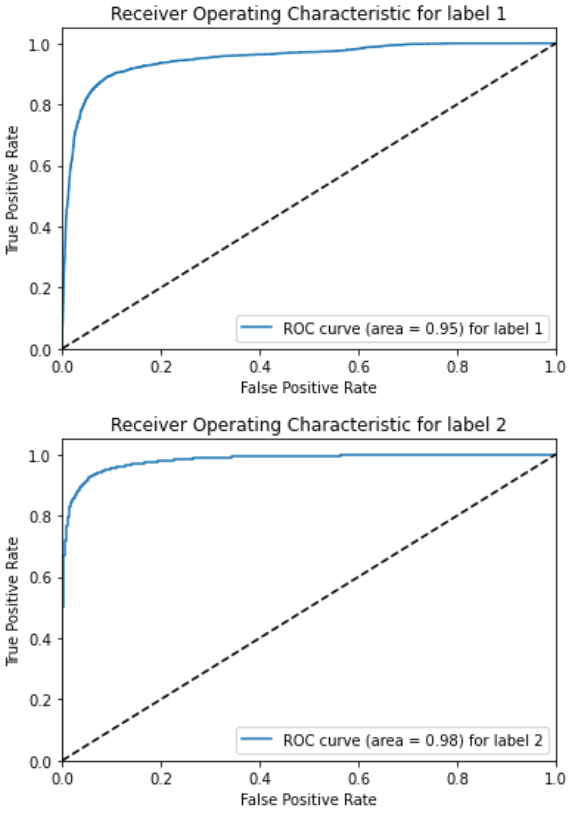
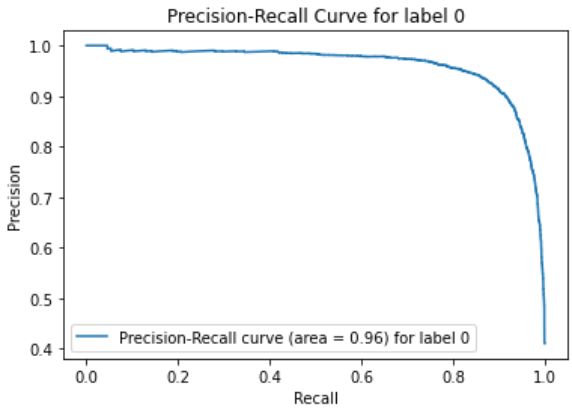


Figure 6 - ROC curves for ELECTRA model prediction on SNLI validation data

The Precision Recall Curves focuses on the positive class and shows the tradeoff between precision and recall for the evaluation set. The curves suggest that the model has the good precision recall for all the labels, with the best performance on Label-3 followed by Label-0 and Label-1.



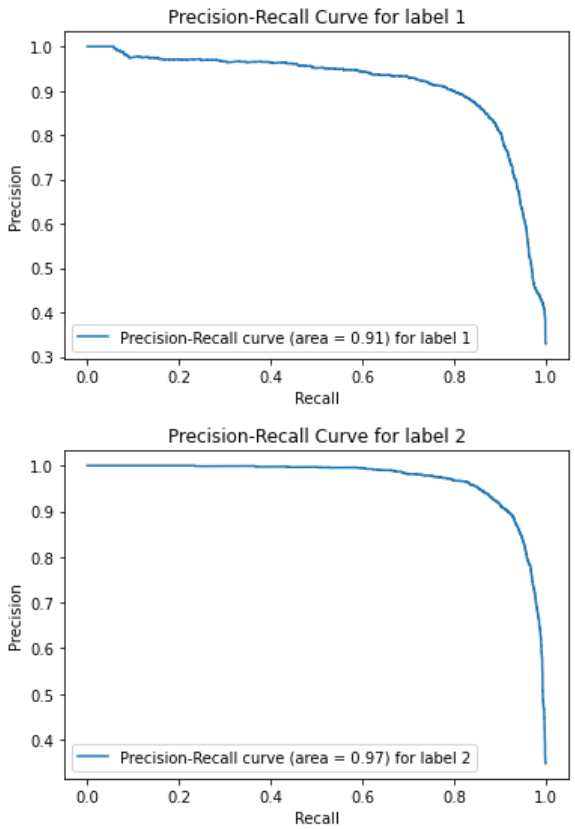


Figure 7 – Precision recall curves for ELECTRA model prediction on SNLI validation data

4.b.(iv) Examples and Classification of Errors:

A. Contradiction predicted as Entailment:

A.1. Unable to recognize contradiction in action predicates:

Premise: Big hairy dog chews on a bone while lying on a furry toy.
Hypothesis: The dog is laying on a pointy object
Actual Label: 2 (Contradiction)
Predicted Label: 0 (Entailment)

Premise: Inside the igloo, the young man gets ready for his overnight stay.
Hypothesis: The young man gets ready to leave.
Actual Label: 2 (Contradiction)
Predicted Label: 0 (Entailment)

A.2 Incorrect attribution of second actor's actions to first actor

Premise: A wrestler in purple curls up in a ball while one in blue celebrates.
Hypothesis: The wrestler in purple celebrates.
Actual Label: 2 (Contradiction)
Predicted Label: 0 (Entailment)

A.3. Interchanging of subject and actions:

Premise: A band playing with fans watching.
Hypothesis: A band watches the fans play
Actual Label: 2 (Contradiction)
Predicted Label: 0 (Entailment)

B. Entailment predicted as Contradiction:

B.1. Failing to spot context that supports entailment:

Premise: A group of people gathered to watch fireworks.
Hypothesis: Everyone is outside looking into the sky.
Actual Label: 0 (Entailment)
Predicted Label: 2 (Contradiction)

B.2. Inability to recognize the change in subject:

Premise: A young boy carries a green, white, and red flag and walks next to a woman.
Hypothesis: The woman is walking next to the boy carrying the flag.
Actual Label: 0 (Entailment)
Predicted Label: 2 (Contradiction)

B.3. Inability to understand paraphrasing:

Premise: A woman is pushing her bike with a baby carriage in front.
Hypothesis: A woman is taking her child for a stroll.
Actual Label: 0 (Entailment)
Predicted Label: 2 (Contradiction)

C. Entailment predicted as Neutral:

C.1. Error due to specificity variation in hypothesis compared to premise:

Premise: A firefighter, in full uniform, looks off into the distance
Hypothesis: This firefighter is on duty.
Actual Label: 0 (Entailment)
Predicted Label: 1 (Neutral)

C.2. Inability to simplify extra details in premise:

Premise: Man in gray zippered jacket and red shirt pets a brown dog and holds a young girl in pink shirt, while a woman next to him holds an infant boy and a gray dog.

Hypothesis: A man holds a small chil
d while petting his dog with his gir
lfriend is holding her son and her d
og.
Actual Label: 0 (Entailment)
Predicted Label: 1 (Neutral)

C.3. Inability to understand synonyms in premise:
Premise: A boy poses in karate form
and uniform.
Hypothesis: A karate student poses i
n his gi
Actual Label: 0 (Entailment)
Predicted Label: 1 (Neutral)

D. Neutral predicted as Entailment
D.1. Oversimplification Error:
Premise: A motorcycle races.
Hypothesis: A bike is going fast.
Actual Label: 1 (Neutral)
Predicted Label: 0 (Entailment)

D.2. Error due to partial match between hypothesis and premise:
Premise: A black dog is drinking nex
t to a brown and white dog that is l
ooking at an orange ball in the lake
, whilst a horse and rider passes be
hind.
Hypothesis: Two short haired dogs of
different colors drinking by a lake.
Actual Label: 1 (Neutral)
Predicted Label: 0 (Entailment)

D.3. Inability to distinguish adverb in action:
Premise: People work diligently on t
here computers at their desks.
Hypothesis: Inside the office buildi
ng several people composed articles
on their computers
Actual Label: 1 (Neutral)
Predicted Label: 0 (Entailment)

E. Neutral predicted as Contradiction
E. 1. Incorrect Gold label:
Premise: Two entertainers on a stage
performing acrobatic moves.
Hypothesis: Two entertainers are per
forming acrobatics in a car.
Actual Label: 1 (Neutral)
Predicted Label: 2 (Contradiction)

E. 2. Inaccuarate inference from action:
Premise: A race car sits in the pits
Hypothesis: The car is being repaire
d.
Actual Label: 1 (Neutral)
Predicted Label: 2 (Contradiction)

E. 3. Unable to relate different types of information like location and action in neutrality.
Premise: A young child brushes his o
r her teeth while holding a remote.
Hypothesis: The young boy is in the
bathroom.
Actual Label: 1 (Neutral)
Predicted Label: 2 (Contradiction)

F. Contradiction predicted as Neutral
F. 1. Inability to distinguish improbable actions by a doer, here from the hypothesis, butterflies in general do not try to kill.
Premise: A butterfly has lit on the
head of this sleeping toddler.
Hypothesis: A butterfly is trying to
kill a sleeping toddler.
Actual Label: 2 (Contradiction)
Predicted Label: 1 (Neutral)

F. 2. Inability to understand time-related sequence of actions:
Premise: A little boy with an orange
shirt is playing outside with bubble
s.
Hypothesis: The boy was given a bubb
le bath after playing outside.
Actual Label: 2 (Contradiction)
Predicted Label: 1 (Neutral)

F. 3 Inability to contradict reversal of roles, between bicyclists and jeep in the case below.
Premise: Three people on bicycles, p
edaling down the side of a road, wit
h a jeep following them.
Hypothesis: Three people on bicycles
are chasing a jeep.
Actual Label: 2 (Contradiction)
Predicted Label: 1 (Neutral)

4.b.(v) Possible rules to identify above challenges:

(i) If the premise and hypothesis have roles reversed, it must be marked as a contradiction.

(ii) If the subject's action after the verb is not matching in the hypothesis and premise, it must be marked as contradiction

(iii) Identifying actions and corresponding actors can help in identifying contradiction
(iv) Improve model's ability to understand synonyms
(v) Improve model's understanding of actions and predicates
(vi) Improve model's ability to compare level of detail available in hypothesis versus premise.
(vii) Model should be able to associate events in a time-wise sequential manner
(viii) Model must not infer contradiction from neutral extensions of a sentence.
(ix) Improve Gold Label accuracy by exposing model to more data.

5 Improving Model Performance

To improve accuracy and address the results from the model based on the findings from the hypothesis-only analysis, we use the approach from the papers [35](Nie et al.,2020) and [36](Williams et al.,2020). In this approach, a new large-scale NLI dataset was composed utilizing a human-and-model in-loop process that acts as adversarial data. The dataset was collected to create a large benchmark dataset that can last long and to generalize models that contain dataset artifacts and create a better model.

The dataset consists of three rounds of data, with increasing complexity and difficulty in classification. The paper [35] (Nie et al.) analyzed round-1's adversarial data using a BERT-Large model [10] (Devlin et al.) that was trained on SNLI and MNLI [34] (Williams et al., 2018) datasets as its base model for training. For round-2, the paper utilized a RoBERTa model [11] (Liu et al., 2019) that was trained on SNLI, MNLI, NLI-version of FEVER [37] (Thorne et al., 2018) along with data from round-1 as its base model. Round-3 used RoBERTa model trained on a wider and more diverse range of data sources along with round-2 data.

The approach we take uses the Google Small Electra model trained on the SNLI dataset as the base model (base). As the premises in the SNLI dataset are taken from captions of images, they are much shorter compared to the ANLI dataset's premises. The ANLI dataset contains longer premises which the paper describes are harder as they are made using longer sentences from multiple sources.

6 Results

We start by analyzing the performance of the Electra model that was trained on SNLI training data (base) on round-1, round-2 and round-3 test data set of ANLI dataset. This is a low performance as the dataset labels are equally distributed.

ANLI Test data	Round-1	Round-2	Round-3
Accuracy	30.3%	32.8%	33.41%
Loss	2.38	2.30	2.22

Table 7 - Evaluation of Base SNLI trained model on ANLI test data

To improve the base model's performance, instead of training each round of the ANLI dataset separately, we concatenate ANLI's round-1, round-2, and round-3 into one single dataset. We use two methods for training the small Electra model:

- **Approach-1 (A1):**
  Train the Electra model that was already trained on SNLI training set on the ANLI concatenated training data.
- **Approach-2 (A2):**
  Train the Electra model from scratch on the training data that is constructed by concatenating training data from SNLI dataset with round-1, round-2 and round-3 training data of ANLI dataset.

In both approaches, the training is done for three epochs on the dataset selected. We do not try to achieve state-of-the-art performance by increasing the number of epochs and fine-tuning because the purpose of this study is to compare the performance difference that occurs with the addition of adversarial data

with approximately the same amount of compute utilization on a basic model compared to the more robust models utilized in the ANLI paper. ANLI dataset is available in Google's Hugging Face [38] (Wolf et al., 2020) implementation library, and is integrated with our code for running the model.

In the first approach, after the SNLI pre-trained Electra model is trained on concatenated ANLI dataset for 3 epochs, it achieves an accuracy of 81.55% accuracy on the SNLI validation dataset, this is a drop from 89.15% accuracy that the model had previously on this set. This could be due to the unlearning of artifacts from the SNLI dataset that the model had learned earlier, due to exposure to the newly presented adversarial information present in ANLI dataset. We aim to improve the generalization of the model in NLI task, so we further check the performance of the approach-1 updated model on the ANLI test data:

ANLI Test data	Round-1	Round-2	Round-3
Accuracy	51.39%	40.9%	43.75%
Loss	1.51	1.57	1.97

Table 8 - Evaluation of ANLI test data on pretrained SNLI model trained on combined ANLI train data

We observe that the model from approach 1 can generalize better based on the increase in 20% accuracy increase on the round-1 ANLI test set and around 10 % increase in the round-2 and round-3 ANLI test sets.

In approach 2, after training the Electra model from scratch for 3 epochs on concatenated ANLI and SNLI datasets, we observe an increase in evaluation accuracy on SNLI validation data from 89.15% to 89.50%.

ANLI Test data	Round-1	Round-2	Round-3
Accuracy	52.39%	43.09%	41.58%
Loss	1.59	2.09	2.13

Table 9 - Evaluation of ANLI test data on ELECTRA base model trained combined SNLI+ANLI train data

When compared to the model in approach 1, the model in approach 2 performs better by 1% accuracy for round-1 and by 3% for round-2 ANLI test data. For round-3 ANLI test data, there is a decrease in accuracy by 2.17%.

7.a. Performance of Base SNLI Electra model versus Approach-1 (A1) model on SNLI validation dataset:

We saw that approach-1 accuracy on the SNLI validation data is less than the base SNLI trained Electra model. Some of the errors that occurred after training on adversarial data are as follows:

- Correct Neutral classification in base model, incorrectly classified as Contradiction:

```
Premise: Two young children in blue
jerseys, one with the number 9 and o
ne with the number 2 are standing on
wooden steps in a bathroom and washi
ng their hands in a sink.
Hypothesis: Two kids at a ballgame w
ash their hands.
Correct Label : 1
Base SNLI prediction : 1
A1 prediction : 2
```

These errors account for 12.8% of errors.

- Correct Neutral classification in base model, incorrectly classified as Entailment:

```
Premise: An asian woman sitting outs
ide an outdoor market stall.
Hypothesis: A woman is selling stuff
in a flea market.
Correct Label : 1
Base SNLI prediction : 1
A1 prediction : 0
```

These errors account for 8.5% of errors.

- Correct contradiction in base model, incorrectly classified as entailment:

```
Premise: Number 13 kicks a soccer ba
ll towards the goal during children'
s soccer game.
Hypothesis: A player fighting in a s
occer game.
Correct Label : 2
Base SNLI prediction : 2
A1 prediction : 0
```

These errors account for 1.7% of errors. This type of error is the minimum among the error s.

- Correct contradiction in base model, incor rectly classified as neutral:

```
Premise: Two young boys of opposing teams play football, while wearing f ull protection uniforms and helmets.
Hypothesis: dog eats out of bowl.
Correct Label : 2
Base SNLI prediction : 2
A1 prediction : 1
```

These errors account for 45.42% of errors. Th is type of error is the maximum among the err ors.

- Correct entailment in base model, incorre ctly classified as contradiction:

```
Premise: Two children in a bag race competing against each other while a nother boy watches on.
Hypothesis: There are three children .
Correct Label : 0
Base SNLI prediction : 0
A1 prediction : 2
```

These errors account for 14.91% of errors.

- Correct entailment in base model, incorre ctly classified as neutral:

```
Premise: The two farmers are working on a piece of John Deere equipment.
Hypothesis: Men are working on John Deere equipment
Correct Label : 0
Base SNLI prediction : 0
A1 prediction : 1
```

These errors account for 16.55% of errors.

7.b. Performance of Base SNLI Electra model versus Approach-2 (A2) model on SNLI validation dataset:

- Incorrect Neutral classification in the base model, correctly classified as Contradiction by A2:

```
Premise: Families waiting in line at an amusement park for their turn to ride.
Hypothesis: People are waiting to se e a movie.
Correct Label : 2
Base SNLI prediction : 1
A2 prediction : 2
```

This type of error correction contributes to 13 .65% of total improvement in accuracy.

- Incorrect Neutral classification in base model, correctly classified as Entailment by A2:

```
Premise: This mother and her daughte r and granddaughter are having car t rouble, and the poor little girl loo ks hot out in the heat.
Hypothesis: The car that belongs to the family is having issues.
Correct Label : 0
Base SNLI prediction : 1
A2 prediction : 0
```

This type of error correction contributes to 14 .67% of total improvement in accuracy.

- Incorrect contradiction in base model, cor rectly classified as entailment by A2:

```
Premise: A man in a bar drinks from a pitcher while a man in a green hat looks on and a woman in a black shir t drink from a glass.
Hypothesis: A women in black drinks.
Correct Label : 0
Base SNLI prediction : 2
A1 prediction : 0
```

This type of error correction contributes to 2. 73% of total improvement in accuracy. This t ype of error correction is the minimum amon g improvement in accuracy.

- Incorrect contradiction in base model, cor rectly classified as neutral by A2:

```
Premise: A shirtless man is singing into a microphone while a woman next to him plays an accordion.
Hypothesis: He is Polish.
Correct Label : 1
Base SNLI prediction : 2
A1 prediction : 1
```

This type of error correction contributes to 37 .54% of total improvement in accuracy. This t ype of error correction is the maximum amon g improvement in accuracy.

- Incorrect entailment in base model, correc tly classified as contradiction by A2:

```
Premise: A woman holds her child out of a red window, next to a Color TV sign.
Hypothesis: "The TV holds a woman wh o holds a child.
```

Correct Label : 2
Base SNLI prediction : 0
A1 prediction : 2

This type of error correction contributes to 10.23% of total improvement in accuracy.

- Incorrect entailment in base model, correctly classified as neutral by A2:

Premise: Two men on bicycles competing in a race.
Hypothesis: Men are riding bicycles on the street.
Correct Label : 1
Base SNLI prediction : 0
A1 prediction : 1

This type of error correction contributes to 21.16% of total improvement in accuracy.

7.c. Performance of Approach-1 (A1) model versus Approach-2 (A2) model on ANLI test data:

For round-1 ANLI test, A2 model performs better by 1% and for round-2, A2 performs better by 3% accuracy. The improvement can be classified as follows:

- Incorrect Neutral classification in the A1 model, correctly classified as Contradiction by A2:

Premise: A trumpet concerto is a concerto for solo trumpet and instrumental ensemble, customarily the orchestra. Such works have been written from the Baroque period, when the solo concerto form was first developed, up through the present day. Some major composers have contributed to the trumpet concerto repertoire, with the best known work being Joseph Haydn's Trumpet Concerto in E-flat.
Hypothesis: A trumpet concerto features multiple trumpets playing a lead line.
Correct Label : 2
A1 prediction : 1
A2 prediction : 2

This type of error correction contributes to 9.17% of total improvement in accuracy.

- Incorrect Neutral classification in A1 model, correctly classified as Entailment by A2:

Premise: The Beijing Municipal Administration & Communication Card (), more commonly known as the Yikatong (literally One-card pass), is a store-value contactless smart card used in Beijing, China, for public transportation and related uses. It is similar to Hong Kong's Octopus card, Singapore's CEPAS, or the Oyster Card used by Transport for London in London, England.
Hypothesis: Beijing Municipal Administration & Communication Card produces smart cards similar to other companies in the same field.
Correct Label : 0
A1 prediction : 1
A2 prediction : 0

This type of error correction contributes to 16.51% of total improvement in accuracy.

- Incorrect contradiction in A1 model, correctly classified as entailment by A2:

Premise: San Andreas is a 2015 American disaster film directed by Brad Peyton and written by Carlton Cuse, with Andre Fabrizio and Jeremy Passmore receiving story credit. The film stars Dwayne Johnson, Carla Gugino, Alexandra Daddario, Ioan Gruffudd, Archie Panjabi, and Paul Giamatti. Its plot centers on an earthquake caused by the San Andreas Fault devastating Los Angeles and the San Francisco Bay Area.", Hypothesis: San Andreas is a American disaster film that takes place on Los Angeles and San Fransisco Bay Area
Correct Label : 0
A1 prediction : 2
A2 prediction : 0

This type of error correction contributes to 13.76% of total improvement in accuracy.

- Incorrect contradiction in A1 model, correctly classified as neutral by A2:

Premise: Susan Sadlowski Garza is a member of the Chicago City Council serving as Alderman for the 10th ward. The 10th ward is located on Chicago's southeast side and includes East Side, Hegewisch, Jeffrey Manor, South Chicago and South Deering. She is serving her first term after defeating Rahm Emanuel ally John Pope in the 2015 election
Hypothesis: Susan Sadlowski Garza never met Rahm Emanuel
Correct Label : 1
A1 prediction : 2
A2 prediction : 1

This type of error correction contributes to 22 % of total improvement in accuracy. This type of error correction is the maximum among improvement in accuracy.

- Incorrect entailment in A1 model, correctly classified as contradiction by A2:

```
Premise: Ghosting refers to the act of breaking off a relationship (often used in the context of intimate relationships) by ceasing all communication and contact with the former partner without any apparent warning or justification, as well as avoiding and/or ignoring and refusing to respond in any way to the former partner \u2019s attempts to reach out or communicate.
Hypothesis: Ghosting is the act of breaking off a relationship but it very often does involve a warning and justification
Correct Label : 2
A1 prediction : 0
A2 prediction : 2
```

This type of error correction contributes to 22.93% of total improvement in accuracy.

- Incorrect entailment in A1 model, correctly classified as neutral by A2:

```
Premise: Big Data is an American electronic music project created by producer, Alan Wilkis. Big Data is best known for its single \"Dangerous\", featuring Joywave, which reached number one on the \"Billboard\" Alternative Songs chart in August 2014, and was certified gold by the RIAA in May 2015.
Hypothesis: Big Data wrote Dangerous with Joywave.
Correct Label : 1
Base SNLI prediction : 0
A1 prediction : 1
```

This type of error correction contributes to 15.59% of total improvement in accuracy.

The reduction in accuracy for the round-3 comparison between A1 and A2 models could be attributed to the higher complexity and difficulty of the hypothesis sentence pairs in this round compared to round-2 and round-1.

7 Discussion and Conclusion

Overall, we can conclude the following from the above results:

- Even though we trained the model by concatenating data of all the rounds together instead of the approach used in the ANLI paper where each round goes through a separate training, there is significant evaluation improvement even with just 3 epochs on the adversarial data.
- We used a simpler model of Electra compared to the stronger models used in the ANLI paper, but we can see that even a simple model can generalize better when trained on adversarial data.
- If we compare the pretrained Electra model trained from scratch with combined SNLI+ANLI data (A2) versus the model trained with ANLI upon the existing SNLI model (A1), both are exposed to the same adversarial data. Even though the performance of training from scratch (A2) seems to be better on the comparatively simple hypothesis-premise pairs from round-1 and round-2, training on pre-existing SNLI Electra model (A1) gives better results in the more complex and difficult examples contained in round-3.
- The better performance of training on pre-existing models could be associated with the model forgetting artifacts learned from older samples that it was trained on as described in paper [39](Yaghoobzadeh et al.2021) and remembering the newer examples that it was exposed to in the latest 3 epochs.
- The compute resources required for training the model on a pre-existing SNLI model (A1) is much lesser than the compute required to train the model from scratch on the combined dataset (A2). This is a big advantage of the A1 model over A2 especially because the accuracy results are similar (and even better that A1 in round-3) .

As we saw in the analysis section, the role of artifacts can be significant depending on the kind of data that is used for training. The availability of adversarial datasets is going to keep

growing as we expose ourselves to newer texts of language. For simpler models like Electra, instead of training the model on combined data from previous pieces of training along with the newly added adversarial data, training on the pre-existing model with just the newly available data is more efficient while also giving better accuracy results of generalization on difficult and complex examples of NLI classification based on this study.

References

[1] Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Martha Palmer, and Valeria de Paiva. 2019. Explaining Simple Natural Language Inference. In Proceedings of the 13th Linguistic Annotation Workshop, pages 132–143, Florence, Italy. Association for Computational Linguistics.

[2] Dagan Ido, Roth Dan, Sammons Mark, Zanzotto Fabio Massimo. Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies. 2013

[3] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets

[4] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online, July. Association for Computational Linguistics.

[5] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

[6] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics

[8] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium, October-November. Association for Computational Linguistics

[9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers

for language understanding. arXiv preprint arXiv:1810.04805.

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

[12] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2020, arXiv:1906.08237.

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942

[14] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Computation and Language.https://doi.org/10.48550/arXiv.2003.10555.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683

[16] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. Available from https://api.semanticscholar.org/CorpusID:49313245.

[17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165

[18] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. In Transactions of the Association for Computational Linguistics, vol. 7, pages 452–466. MIT Press, Cambridge, MA. DOI: 10.1162/tacl_a_00276.

[19] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China, November. Association for Computational Linguistics

[20] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. Transactions of the Association for Computational Linguistics, 8:662–678.

[21] Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In Proceedings of the 56th Annual Meeting of the Association for 6 Computational Linguistics (Volume 2: Short Papers), pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

[22] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy, July. Association for Computational Linguistics.

[23] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. DOI: 10.18653/v1/P18-1128.

[24] Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[25] Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5010–5015, Brussels, Belgium, October-November. Association for Computational Linguistics.

[26] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2023

[27] Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8759–8771, Online, July. Association for Computational Linguistics.

[28] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

[29] Dheeru Dua, Pradeep Dasigi, Sameer Singh, and Matt Gardner. 2021. Learning with instance bundles for reading comprehension.

[30] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9275–9293, Online, November. Association for Computational Linguistics.

[31] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Workshop on Ethics in NLP.

[32] Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo¨ıc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532– 1543

[34] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

[35] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv preprint arXiv:1910.14599.

[36] Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the Adversarial Natural Language Inference Dataset. arXiv preprint arXiv:2010.12729.

[37] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. arXiv preprint arXiv:1803.05355.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al.. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics

[39] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3319–3332, Online, April. Association for Computational Linguistics.