# CS378 NLP Final Project

Bruno Fazzani, Justin Lee
University of Texas at Austin

**Abstract**

Having robust and large datasets is very important for the performance of current natural language processing techniques, with data-intensive neural network based approaches being exceedingly popular. One challenge faced is that of dataset artifacts, implicit features of a dataset that can be learned by language models without actually demonstrating understanding in the language task. In this project, we analyze a popular dataset for the natural language inference task to try and identify and rectify dataset artifacts that can cause models to learn superficial correlations. We use challenge sets to expose specific artifacts and fine-tune our language model by training on these challenge sets, showing greatly improved performance on the subset of examples that are impacted by the dataset artifacts, while also not significantly impacting the performance on the base benchmark dataset.

## 1 Introduction

Natural Language Inference (NLI) is a task in Natural Language Processing to determine the validity of a *hypothesis* given a *premise*. The hypothesis and premise are two separate sentences provided to a model which must then decide if the hypothesis is true (entailment), the hypothesis is false (contradiction), or if the validity of the hypothesis cannot be determined (neutral) given the premise.

One challenge with using modern neural network based approaches is that they are relatively model agnostic, and will learn any available patterns in the training data regardless of prior assumptions on important features and patterns of the data. This can lead to bad generalization outcomes if the dataset used is of poor quality, or contains *dataset artifacts*, which are correlations and patterns in dataset that are external to the training task. This can come from anything from poor distribution of examples to poor construction of training examples.

This is an important issue for natural language processing because many of the currently popular models are transformer-based neural networks that are very data-intensive, so datasets are often too large to be manually examined for any potential issues, as well as simply the origin of the data can greatly influence the properties of the language data. Therefore we are interested in automated approaches to identifying and fixing dataset artifacts to improve the performance of our natural language models.

## 2 Dataset Artifacts

For the NLI task we used the Stanford NLI dataset (Bowman, Angeli, Potts, & Manning, 2015), which contains 570k English sentence pairs formed from image captioning. Each training example consists of two sentences, the premise and the hypothesis, as well as the gold label that describes whether the sentences are related via entailment, contradiction, or are neutral to each other. The gold label is determined as the majority label of 5 annotators, or when there isn't a majority, the gold label is omitted, represented as -1. This occurred in only 785 of the training examples. However, the NLI dataset has been critiqued for its simplicity and lack of variability (Glockner, Shwartz, & Goldberg, 2018). These issues have caused the dataset to contain dataset artifacts which can weaken the overall strength of the Stanford dataset in NLI training. However, it is still a very large and easily accessible dataset due to its public availability and simple format, so we are interested in continuing to use the Stanford dataset and instead explore supplementary means to mitigate and remove dataset artifacts.

In analyzing the SNLI dataset, we identified two dataset artifacts to explore further: word overlap and negations. Word overlap is when both the premise and hypothesis only have small variations between each other, and mostly use the same words. In SNLI, high word overlap shows a propensity for the sentences to be related via entailment. This likely comes from weak or lazy training examples where neutral or contradiction examples are more easily generated from discussing two completely distinct topics, whereas entailment examples can be generated by making small modifications to a sentence that don't influence

the syntax nor semantics of the sentence greatly. In our analysis, we fitted a multiclass logistic regression on the word overlap between the premise and the hypothesis (insert latex equation showing it) to predict the gold label of the example based solely on this metric.

$$word\ overlap = \frac{\#\ of\ shared\ words\ between\ s_1 and s_2}{max(|s_1|, |s_2|)}$$

Our classifier demonstrated a 0.398 mean accuracy on the SNLI test set after being trained on the main training set. This accuracy being higher than the expected for random guessing shows that the word overlap metric has correlations to the gold label, which aren't related to the NLI task itself.

The other dataset artifact is the presence of negating words in the sentence pair. Another simple way of creating contradiction examples would be to take a base sentence, and then simply negate some aspect of it, such as with *The hat is blue* and *The hat is not blue*. The conditional probability of a contradiction gold label given the presence of the negating words not or no is about 40%, as opposed to roughly 30% for the other labels. This shows another spurious correlation unrelated to the NLI task.

| Label | P(Label \| negation) |
|---|---|
| Entailment | 0.294 |
| Neutral | 0.307 |
| Contradiction | 0.399 |

# 3 Challenge Sets and Stress Testing

We analyzed our model by evaluating its performance on adversarial datasets. Adversarial datasets are designed specifically for the purpose of challenging models and testing their performance. These datasets may focus on a specific weakness that a model is likely to have or modify pre-existing datasets to introduce new complexities.

The first dataset we challenged our model with was the Breaking NLI dataset (Glockner et al., 2018). This dataset takes premises from SNLI and replaces a single word with a synonym, antonym, or hypernym to generate hypotheses. For example, from the premise "A little girl is very sad.", a corresponding hypothesis "A little girl is very unhappy" is generated by replacing "sad" with a synonym. This ultimately requires a model to have simple lexical knowledge to make accurate predictions on this dataset. According to Glockner, a model trained on just SNLI should perform significantly worse on Breaking NLI. In the paper, models that achieve 84.7% - 87.9% accuracy on SNLI only got 51.9% - 65.6% on Breaking NLI.

Stress Test for NLI (Naik, Ravichander, Sadeh, Rosé, & Neubig, 2018) analyzed the performance of different models trained on the MultiNLI dataset, categorized the most common errors that the models made, and generated adversarial datasets to target these weaknesses. The categories are generalized into six different language phenomena: word overlap, negation, antonyms, numerical reasoning, length mismatch, and spelling errors. Models often wrongly predicted entailment when a sentence pair had unrelated meanings but a large word overlap. They also tended to predict contradiction in the presence of a negation word like "no" or "not". They in turn failed to predict contradiction for sentence pairs with antonyms simply because of the lack of a negation word. Models had difficulty reasoning with numbers or quantifiers, and they had trouble sorting through noise in the case of a length mismatch or in the presence of spelling errors. Based on this analysis, original data was generated or the MultiNLI dataset was modified to create adversarial datasets for each of the categories. Because we are working with the SNLI dataset, and the stress tests were modeled after the MultiNLI dataset, we aggregated genre matched and genre mismatched data into a single set. Out of the six datasets, we chose the negation, word overlap, numerical analysis, and antonym sets to challenge our model.

# 4 Fine-tuning with Inoculation

For our baseline model we used the ELECTRA-small model, which is a variant of the BERT architecture with a less computationally demanding training method. We trained our base model on 3 epochs of the SNLI training set. After training the base model, we trained multiple fine-tuned models using the challenge sets. For each challenge set except the antonym dataset, we began with the base model and trained an additional 3 epochs on roughly 10% of that challenge set. From the antonym dataset, we fine tuned our model with 10 individual data points and only trained on 1 epoch.

# 5 Results

For each of the datasets, we evaluated our ELECTRA-small model's performance based on accuracy. As expected, the baseline model achieved a high accuracy on SNLI and performed significantly worse on each of the adversarial datasets. The one expectation of this performance was on the Breaking NLI dataset, on which our model achieved an even higher accuracy than on SNLI. This is likely because our model outperforms those that are cited in the paper and is able to overcome the challenge that is posed by Breaking NLI. Because of this result, we decided to not test inoculation for this dataset.

Then, we evaluated the fine-tuned models against their respective datasets. Inoculation proved effective as we saw significant increases in accuracy across each of the stress tests. The one abnormal dataset was the antonyms set. With as few as 10 inoculation data points and a single epoch of training, we saw an increase of 26% in accuracy. This could be because the antonym dataset is smaller than the others or because the task is easier for the model to learn.

Finally, we evaluated the performance of each of our fine tuned models against SNLI to check if inoculation improved or worsened the baseline model's accuracy. Inoculation made no significant change to the overall performance of the baseline model. The very slight decrease in the accuracy is likely because the model is unable to depend on training artifacts like it did before fine tuning.

| Challenge Set | Initial Eval. | Post-inoculation Eval. | Post-inoculation SNLI Eval. |
|---|---|---|---|
| Stanford NLI | 0.891 | - | - |
| Breaking NLI | 0.94 | - | - |
| Negation | 0.46 | 0.73 | 0.887 |
| Word Overlap | 0.62 | 0.74 | 0.886 |
| Numerical | 0.22 | 0.53 | 0.887 |
| Antonyms | 0.65 | 0.91 | 0.887 |

Table 1: Evaluation results

Fine-tuning via inoculation was very successful in improving our models on the challenge sets, significantly increasing the accuracy of the models with very few additional training examples. These fine-tuned models should mitigate some of the problems of the dataset artifacts present in the SNLI dataset.

# References


Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *CoRR*, *abs/1508.05326*. Retrieved from http://arxiv.org/abs/1508.05326

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. *CoRR*, *abs/1805.02266*. Retrieved from http://arxiv.org/abs/1805.02266

Naik, A., Ravichander, A., Sadeh, N. M., Rosé, C. P., & Neubig, G. (2018). Stress test evaluation for natural language inference. *CoRR*, *abs/1806.00692*. Retrieved from http://arxiv.org/abs/1806.00692