# Improving Lexical Reasoning of Natural Language Models via Data Augmentation and Adversarial Training

**Luis Smith**
Department of Computer Science
University of Texas at Austin
luis.s@utexas.edu

## Abstract

Natural Language Inference (NLI) is a common task in Natural Language Processing (NLP) that attempts to predict the relationship between a premise text and a hypothesis text. The field has seen considerable progress in performance across two major datasets, the Stanford NLI Corpus and the Multi-Genre NLI corpus. However, even the best performing models on these datasets fail to have strong lexical reasoning skills, as evidenced by the difficulty of recognizing the relationship between antonyms and contradictions. Our paper seeks to remedy this problem using two different approaches: data augmentation and adversarial training. Adversarial training proves to be the better solution, resulting in a model with better lexical reasoning skills and generalization properties.

## 1 Introduction

Natural Language Processing (NLP) is a growing field within machine learning. One of the canonical tasks in NLP is Natural Language Inference (NLI), which is the task of defining the semantic relation between a premise $p$ and a hypothesis $h$. The hypothesis $h$ is said to either: a) entail, b) contradict, or c) be neutral to the premise $p$ (Kalouli et al., 2019).

The premise $p$ is taken to entail hypothesis $h$ when a human reading $p$ would infer that it is most probably true (Kalouli et al., 2019). Humans labelers are tasked with creating labels for datasets (i.e. "entailment", "contradiction", or "neutral") based on the relationship between the premise and the hypothesis. The goal in NLI is to create natural language models that are able to accurately predict the correct label that describes this relationship.

In our paper, we create a baseline model by using a pre-trained model, ELECTRA-small (Clark et al., 2020), and fine-tuning it on the Multi-Genre Natural Language Inference dataset (Williams et al., 2018). We use an adversarial dataset (Glockner et al., 2018) to show our baseline model underperforms when performing lexical reasoning on antonyms. We then create a data augmentation strategy that adds additional training examples to our training set by modifying the original examples with antonyms and synonyms. With data augmentation, lexical reasoning on antonyms is improved, however generalization is harmed. Our second strategy is adversarial training on the Adversarial NLI (ANLI) dataset (Nie et al., 2020), which contains a significant percentage of examples that require complex lexical reasoning. This proves to be the best strategy, as lexical reasoning on antonyms and generalizations are both significantly improved.

## 2 The Multi-Genre Natural Language Inference Dataset

One of the premier datasets created for the task of NLI is the Multi-Genre Natural Language Inference corpus. Henceforth, we will refer to this dataset as MNLI. At 433k examples, this resource is one of the largest corpora available for NLI (Williams et al., 2018). It also improved upon the only large human-annotated corpus for NLI at the time, the Stanford NLI Corpus (SNLI, Bowman et al., 2015), by increasing the level of coverage and difficulty. As MNLI offers ten distinct genres of written and spoken English, it made it possible to evaluate systems on nearly the full complexity of the language (Williams et al., 2018).

## 3 Our Baseline Model

State of the art performance on NLI tasks are achieved by large pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and other transformer-based models. The pre-trained language models are further fine-tuned by training on downstream tasks (Devlin et al., 2019).

For our baseline pre-trained model we chose the ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)-small model. The main benefit of the ELECTRA-small model is its computation efficiency. Given the same model size, data, and compute, ELECTRA substantially outperforms masked language modeling (MLM) based methods such as BERT and XLNet (Clark et al., 2020).

The ELECTRA-small model was then fine-tuned by training on the MNLI dataset. Henceforth, we will refer to this fine-tuned model as ELECTRA-MNLI. The baseline performance of ELECTRA-MNLI when evaluated across the test sets in the SNLI and MNLI dataset can be seen in Table 1. We include the SNLI test set to see how well the model is generalizing to different test sets. Overall, the model performs best on entailment and contradiction labels, whereas it struggles more with neutral labels.

Table 1: ELECTRA-MNLI Accuracy Across Different Test Sets

| Label | SNLI | MultiNLI |
|---|---|---|
| Entailment | 83.1% | 83.3% |
| Neutral | 69.0% | 77.6% |
| Contradiction | 77.9% | 84.6% |
| Overall | 76.8% | 81.9% |

## 4 Breaking our Baseline Model with Antonyms

Despite ELECTRA-MNLI's reasonable performance on test datasets, there are some doubts as to what exactly the model is actually learning. A research paper (Glockner et al., 2018) noted that many state of the art models at the time lacked generalization ability, and failed to capture many simple inferences that require lexical and world knowledge. The authors proved this by creating a simple adversarial test set with premises taken from the SNLI training set. For each premise, several hypotheses were generated by replacing a single word within the premise by a different word. Entailment examples were generated by replacing a word with its synonym or hypernym, while contradiction examples were created by replacing words with mutually exclusive co-hyponyms and antonyms. The paper suggested that these adversarial examples "broke" the existing models. Henceforth, we will refer to this adversarial test set as the BNLI test set.

At the time of publication, the best performing model on the BNLI test set was KIM, with an accuracy of 83.5% (Glockner et al., 2018). KIM was a neural model that incorporated external lexical knowledge (from WordNet) (Glockner et al., 2018). However, large transformer-based pre-trained models such as BERT and ELECTRA, did not yet exist. Since large pre-trained models are now state-of-the-art, would these models also break under the BNLI test set?

We investigated by looking at the overall performance of the ELECTRA-MNLI model on the BNLI test set. We can see the performance across different labels in Table 2.

Table 2: ELECTRA-MNLI Accuracy Across BNLI Test Set

| Label | Count | Accuracy |
|---|---|---|
| Entailment | 982 | 98.6% |
| Contradiction | 7164 | 92.2% |
| Neutral | 47 | 17.0% |
| Overall | 8193 | 92.5% |

The overall accuracy of ELECTRA-MNLI at 92.5% performs much better than the best overall accuracy at the time of publication, which was KIM at 83.5%. However, it's important to understand how the ELECTRA-MNLI is performing across the different categories of examples. Are there still examples that the model struggles with? We look at Table 3 for the model's accuracy across different categories.

The ELECTRA-MNLI model seems to have a nearly perfect grasp on synonym substitution, with 100.0% accuracy in the "Synonyms" category. This strong performance may be due to the power of the learned embeddings from the pre-training process on a large language corpus.

However, we see that the ELECTRA-MNLI model struggles much more with antonym substition, achieving an accuracy of 92.2% in the "Antonyms" category and 79.2% in the "Antonyms_Wordnet" category. We refer to Table 4 for examples where the ELECTRA-MNLI model fails to predict antonyms correctly. Humans generally would not have trouble predicting the correct label of contradiction for these examples. This calls into question whether the ELECTRA-MNLI

Table 3: ELECTRA-MNLI Accuracy on BNLI Test Set Categories

| Category | Count | Accuracy |
|---|---|---|
| **Synonyms** | **894** | **100.0%** |
| Countries | 613 | 98.7% |
| Cardinals | 759 | 98.6% |
| Ordinals | 663 | 97.1% |
| Nationalities | 755 | 96.8% |
| Materials | 397 | 96.7% |
| Colors | 699 | 94.1% |
| Instruments | 65 | 92.3% |
| **Antonyms** | **1147** | **92.2%** |
| Planets | 60 | 90.0% |
| Rooms | 595 | 89.6% |
| Drinks | 731 | 82.1% |
| **Antonyms_Wordnet** | **706** | **79.2%** |
| Vegetables | 109 | 49.5% |

model can perform basic lexical reasoning to infer that the same scenario, but with an antonym, generally should lead to a contradiction.

Table 4: BNLI Antonym Examples

| Premise | Hypothesis | Label | Predicted Label |
|---|---|---|---|
| A girl jumping into a pool, with a man **standing** near. | A girl jumping into a pool, with a man **running** near. | Contradiction | Entailment |
| A man in a brown shirt is holding a hat that he just **bought**. | A man in a brown shirt is holding a hat that he just **sold**. | Contradiction | Entailment |

Next, we explore various techniques to try to improve ELECTRA-MNLI's performance on antonyms, while attempting to ensure generalization across various test sets.

## 5 Can Data Augmentation Help?

Data augmentation refers to strategies for increasing the diversity of training examples without explicitly collecting new data. Most strategies either add slightly modified copies of existing data or create synthetic data, aiming for the augmented data to act as a regularizer and reduce overfitting when training ML models (Feng et al., 2021).

The original MNLI dataset may not have enough examples of antonyms in the premise and the hypothesis, and therefore may not be able to explicitly learn when an antonym would imply a contradiction. Thus, our approach is to generate additional training examples by modifying training data from the original MNLI dataset.

Our initial approach involved the use of the "nlpaug" python library (Makcedward), which allows us to generate synthetic data without manual effort. Using the "AntonymAug" function, which finds antonyms based on WordNet (Makcedward), we were able to take unique premises from the MNLI data, and create hypotheses that would result in a contradiction of the premise. Our approach is to replace one word from each hypothesis. We removed any examples where the "AntonymAug" function failed to find an antonym, and thus produced a hypothesis identical to the premise. Examples of augmented training examples that use antonym substitution to create a contradiction are shown in Table 5.

Table 5: Data Augmentation using MNLI Dataset: Contradiction via Antonyms

| Premise | Hypothesis | Label |
|---|---|---|
| The **other** men shuffled. | The **same** men shuffled. | Contradiction |
| Postal Service were to **reduce** delivery frequency. | Postal Service were to **expand** delivery frequency. | Contradiction |

To test our approach, we used the same ELECTRA-small pre-trained model, but now fine-tuned the model on all of the original training examples plus our augmented examples. We refer to this model as MNLI+AUGMENT_ANTONYM, without referring to ELECTRA for brevity. In Table 6, we show the relative performance of this model vs. our baseline ELECTRA-MNLI model. We show comparisons for Antonyms and also include all of the test datasets discussed thus far to see how well the model is generalizing.

We see that that performance greatly im-

Table 6: Relative Accuracy of MNLI+AUGMENT_ANTONYM vs. ELECTRA-MNLI

| Dataset or Data Slice | Relative Accuracy |
|---|---|
| MNLI | -0.4% |
| BNLI | -2.4% |
| SNLI | -0.1% |
| BNLI - Antonyms | 7.7% |
| BNLI - Antonyms (WordNet) | 16.3% |

proved on Antonyms. However, this came at the expense of generalization, as the MNLI+AUGMENT_ANTONYM model performed worse on the MNLI, BNLI, and SNLI datasets.

We speculated the underperformance of MNLI+AUGMENT_ANTONYM may have been due to an imbalanced training dataset. The original MNLI training set has equal splits between entailment, neutral, and contradiction. Since we were nearly doubling the number of contradictions, the dataset would now be imbalanced. High-class imbalance, often observed in big data, makes identification of the minority class by a model complex and challenging because a high-class imbalance introduces a bias in favor of the majority class (Leevy et al., 2018)

In order to test this theory, we attempted to use data augmentation to generate new "entailment" and "neutral" examples to balance the dataset. To generate entailment examples, we took existing sets of premises and hypotheses that had entailment labels. We generated additional entailment examples by keeping the premise the same, and keeping the hypothesis the same except for replacing one word by its synonym. We followed the same process for existing sets of premises and hypotheses that had neutral labels to generate additional neutral examples. Examples of original and augmented training examples can be found in Table 7.

To test our approach, we used the same ELECTRA-small pre-trained model, but now fine-tuned the model on all of the original training examples plus our new set of augmented examples. We refer to this model as MNLI+AUGMENT_ALL. In Table 8, we show the relative performance of this model vs. our baseline ELECTRA-MNLI model. The results show that this new data augmentation process did not improve overall performance, as we underperform across all metrics vs. the MNLI+AUGMENT_ANTONYM model.

Table 7: Data Augmentation using MNLI Dataset: Synonym Substitution

| Premise | Hypothesis | Label | Example Type |
|---|---|---|---|
| and it is nice talking to you all righty | I **talk** to you every day. | Neutral | Original |
| and it is nice talking to you all righty | I **spill** to you every day. | Neutral | Augmented |
| We stink all the time. | We always **stink**. | Entailment | Original |
| We stink all the time. | We always **reek**. | Entailment | Augmented |

Table 8: Relative Accuracy of MNLI+AUGMENT_ANTONYM vs. ELECTRA-MNLI

| Dataset or Data Slice | Relative Accuracy |
|---|---|
| MNLI | -0.2% |
| BNLI | -3.0% |
| SNLI | -1.6% |
| Antonyms | 7.1% |
| Antonyms (WordNet) | 15.0% |

In order to understand why this process may have failed to generalize, we took a closer look at some of the examples generated by data augmentation. We found that certain examples had issues, such as being nonsensical or ungrammatical. We can see examples of these problematic examples in Table 9.

To fix this, we would ideally need to ensure that augmented examples replaced existing words with contextualized synonyms and antonyms to help ensure that the augmented hypotheses make sense.

Table 9: Problematic Augmented Examples

| Premise | Hypothesis | Example Type | Annotation |
|---|---|---|---|
| Were they in there? | Were they **supposed** to be in there? | Original | |
| Were they in there? | Were they **speculate** to be in there? | Augmented | "Speculate" is a synonym for "supposed" but does not make sense here. |
| 'Uh, hi guys,' Daniel said meekly. | Daniel **greeted** everyone even though he was shy. | Original | |
| 'Uh, hi guys,' Daniel said meekly. | Daniel **greet** everyone even though he was shy. | Augmented | Replacing "greeted" with "greet" makes this sentence ungrammatical. |
| i'm **going** to be able to make my own clothes | i'm **stay in place** to be able to make my own clothes | Augmented | "Stay in place" is an antonym for "going" but does not make sense here. |

To replace with a WordNet Synonym and Antonym with the same part of speech as the original word does not seem to be sufficient. Alternatively, we would need a human reader to audit the augmented examples and discard any that did not make sense. Since we currently do not have the resources to manually audit the data, we did not find data augmentation to be a viable approach for improving lexical reasoning about antonyms, while maintaining generalization across various test data sets.

## 6 Adversarial Training

To improve the robustness of a natural language model, one method is adversarial training, where the model is trained on both original examples and adversarial examples. Adversarial training can be viewed as a data augmentation method where hard examples are added to the training set (Yoo and Qi, 2021).

In order to improve lexical understanding and to generalize better, our baseline model would benefit from more challenging training examples. The Adversarial NLI (ANLI) dataset was created to be more difficult than previous NLI datasets, by design. In ANLI, human annotators were employed to act as adversaries, and encouraged to find vulnerabilities that fool the model into misclassifying, but that another person would correctly classify (Nie et al., 2020). In a deep dive analysis of the ANLI dataset, it was discovered that 21.2% of examples contained lexical reasoning (Williams et al., 2020). The definition of lexical reasoning was LEXICAL-DISSIMILAR (Antonymy, Contrast) or LEXICAL-SIMILAR (Overlap, Similar) reasoning (Williams et al., 2020).

Since our goal is to improve lexical reasoning on antonyms, we conjectured that including this dataset would lead to an improvement in overall performance on the test sets. For this reason, we trained an ELECTRA-small model and finetuned it on an enriched dataset that used our original MNLI training examples plus the ANLI training examples. We used all 3 rounds of training examples from ANLI (Nie et al., 2020). We refer to this model as MNLI+ANLI. The results can be seen in Table 10.

Table 10: Relative Accuracy of MNLI+ANLI vs. ELECTRA-MNLI

| Dataset or Data Slice | Relative Accuracy |
|---|---|
| MNLI | 0.4% |
| BNLI | 2.1% |
| SNLI | 0.5% |
| BNLI - Antonyms | 3.3% |
| BNLI - Antonyms (WordNet) | 4.0% |

We can see that there was an improvement of 3.3% on Antonyms and 4.0% on Antonyms (WordNet) in our BNLI test set. We also improved our generalization capabilities, showing a substantial improvement of 2.1% on the BNLI test set, and

modest improvements of 0.4% on the MNLI test set and 0.5% on the SNLI test set.

## 7 Conclusion

The BNLI dataset was one of the early breakthrough datasets that exposed issues with language models' ability to do natural language inference. Although a modern pre-trained language model such as ELECTRA performs well on this dataset, the model still struggles with basic lexical reasoning on antonyms.

We explored two techniques to try to improve overall performance. The first technique was data augmentation that sought to produce additional examples using antonyms and synonyms. This method was successful in improving reasoning about antonyms, but harmed overall generalization capabilities. We conjectured this was due to the poor quality of examples generated via the data augmentation process.

The second method we attempted was to supplement our training data with an adversarial dataset. The goal was to provide more challenging examples that involved more advanced lexical reasoning. Even though the ANLI training data comes from a different distribution than the original MNLI data, it did not hurt model performance. On the contrary, the complex lexical reasoning in the training examples improved performance on our desired subset of data, antonyms, while also generalizing better than our baseline ELECTRA-MNLI model. Thus, we can see the benefit of training on relevant adversarial datasets in order to create language models with improved lexical understanding.

## References


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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

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 Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

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.

Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A. Bauder, and Naeem Seliya. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1).

Makcedward. Makcedward/nlpaug: Data augmentation for nlp.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. Anlizing the adversarial natural language inference dataset.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic. Association for Computational Linguistics.