Challenging the Limits of NLP: Addressing Adversarial Vulnerabilities in Language Models

Abstract

It has come to light in recent years that state-of-the-art Natural Language Processing (NLP) models can achieve incredible performance metrics across a variety of tasks. However, we have begun to question what these models are truly learning as a result of them not being able to perform well on challenging examples such as adversarial datasets which are created with the purpose of exploiting vulnerabilities in the model. In order for NLP models to have true, human-level reading comprehension skills, it needs to be able to decipher examples that may be infrequent, but are present in normal linguistic interactions. An example of such is sarcasm which is covered later in the paper. We will explore the limitations of current NLP models such as Electra and its inability to achieve high performance on adversarial examples. Furthermore, we will demonstrate improvements on model performance through further training on adversarial examples as well as improvements on specific adversarial tasks.

1 Introduction

With the goal of creating more robust Natural Language Processing models that are better suited to understanding the various intricacies of the human language, we need to have an understanding of how these pretrained models are built and why they are incapable of understanding certain examples. Many state-of-the-art models that are frequently used for NLP tasks are trained in a way where the model is able to learn to maximize performance on the test set without truly gaining a deeper semantic understanding. Without seeing examples of the human language that are less frequent but more complex, the models are unable to show a true mastery of reasoning capabilities. For this reason, we choose to evaluate our baseline model on adversarial challenge sets which contain intentionally misleading examples that a human would understand but the model may not. For this work, we are using the ELECTRA-small (Clark et al., 2020) model and training it upon The Stanford NLI dataset (Bowman et al., 2015). This dataset was chosen as we focus on the natural language inference task (NLI) in which the model is presented with a premise and a hypothesis. Based on the premise, the model then decides and classifies the hypothesis as an entailment, a neutral statement, or a contradiction. Entailments are when the hypothesis logically follows the premise. Neutral is when the hypothesis cannot be guaranteed by the premise. Contradictions occur when the hypothesis cannot be true given the premise. The model outputs “0” for entailment, “1” for neutral, and “2” for contradiction. This setup allows us to evaluate how well the model is assessing the relationship between premise and hypothesis, thus establishing a baseline for reading comprehension.

Premise	Two women are embracing while holding to go packages.
Hypothesis	Two woman are holding packages.
Label	0 (Entailment)

Table 1: An example from the Stanford NLI dataset. Hypothesis can be determined true based on the premise, thus the model outputs label “0” for entailment.

Most of the examples in the Stanford NLI dataset are straightforward and lack aspects of ambiguity or confusion. In many natural language situations, changing one word in a sentence can vastly alter the overall meaning or sentiment. We pose the question: Is the model truly learning the meaning of the premise or is it learning something else such as the overlap in words between the hypothesis and premise to predict the label? Testing our baseline model on adversarial data gives us insight into this question. Instead of creating our own adversarial challenge set, we have taken one that has already had research performed on it, the ANLI dataset (Nie et al., ACL 2020).

Premise	The Great Mall of the Bay Area (often simply called The Great Mall) is a large indoor outlet shopping mall in Milpitas, California built by Ford Motor Land Development and Petrie Dierman Kughn in 1994. It was acquired by Mills Corporation in 2003, and by the Simon Property Group in April 2007. The mall contains approximately 1.4 million square feet of gross leaseable area.
Hypothesis	The Great Mall was bought by the Simon Property Group in 2003
Label	2 (Contradiction)

Table 2: Example from ANLI dataset (Nie et al., ACL 2020). The reason this is considered adversarial is because the model may determine this to be an entailment as the Great Mall was purchased by the Simon Property, but not in 2003. The token overlap between the hypothesis and premise may cause the model to output entailment.

Through our experiments, we will demonstrate that models such as Electra have gaps in textual understanding that cause it to perform poorly on examples that may be slightly confusing. When evaluating the baseline model on our adversarial set, we will see that accuracy is relatively low at 29.9% in comparison to 89.3% accuracy when evaluated against the original Stanford NLI evaluation set. Our experiments will prove that these models can be made more robust through further training on specific adversarial examples. We show that feeding the model more adversarial examples will

increase accuracy by nearly 20% while decreasing accuracy on the original evaluation set only slightly. Furthermore, we will validate this method of training on challenge sets by creating our own dataset containing examples of sarcasm, which is a specific subset of adversarial examples.

2 Methods

2.1 Error Analysis

Our first task was to demonstrate that modern Natural Language Processing models are unable to pick up on certain nuances in the human language, thus proving a lack of true reading comprehension skills. We began by training the pretrained model, ELECTRA-small (Clark et al., 2020), on the The Stanford NLI dataset (Bowman et al., 2015) so that the model could perform natural language inference tasks of predicting premise-hypothesis relationships. For context, the Electra Small model is a scaled down version of the full Electra model. While scaled down, it still provides high performance and is computationally more efficient. The Stanford NLI dataset (SNLI) was created with the purpose of allowing ML models to learn textual entailment relationships. After training upon this data and testing on an unseen evaluation set, we were able to see that the model was able to correctly identify 89.3% of unseen premise-hypothesis pairs in the evaluation set containing nearly 10,000 examples.

Now with the understanding that the model is able to perform well under simple examples, we want to demonstrate its lack of comprehension by evaluating it on the adversarial examples. After passing the model 1000 pairs from the ANLI dataset (Nie et al., ACL 2020) and performing inference, the model was only able to achieve an accuracy of 29.9%, thus implying that the model has shortcomings when it comes to learning the true meaning behind each premise.

Our first thought was to check the distribution of predicted labels to make sure that our model was not biased towards one class or another. We determined there was no class imbalance as the distribution of labels was relatively even for all pairs. The model was not getting any particular label incorrect more frequently than other labels as the distribution of incorrect labels was also even. While the model makes many errors, the most common type of error that we identified was for pairs where the ground truth label was supposed to be contradiction, but the model predicted entailment. This occurred more often than a prediction of neutral. For the incorrect predictions with ground truth entailment or neutral, the errors were spread evenly across the other two labels. This implies that for this adversarial evaluation set, the model when wrong is more likely to classify a contradiction as entailment rather than neutral. Therefore we can determine that the model is struggling to identify conflict between the premise and hypothesis.

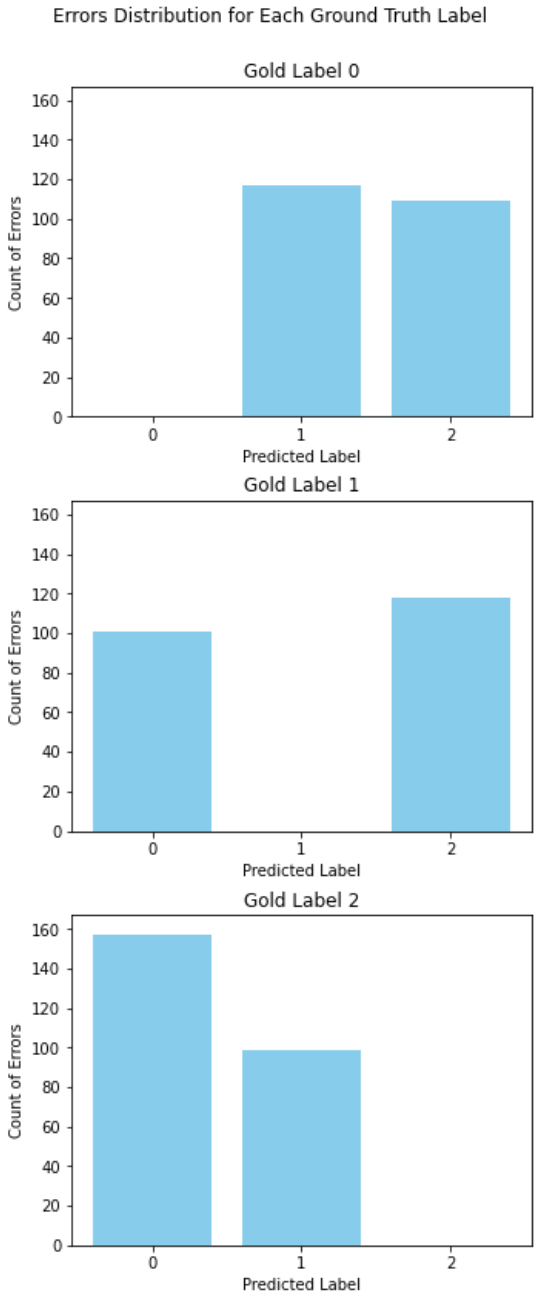

Figure 1: When the ground truth is contradiction, the model more often predicts entailment as opposed to neutral. When the ground truth is entailment or neutral, the errors are evenly distributed.

An example of this specific behavior can be seen in the next table where the model predicts entailment as it sees a major overlap in tokens between the hypothesis and premise even though the hypothesis is false.

Premise	Takaaki Kajita (梶田 隆章 , Kajita Takaaki) is a Japanese physicist, known for neutrino experiments at the Kamiokande and its successor, Super-Kamiokande. In 2015, he was awarded the Nobel Prize in Physics jointly with Canadian physicist Arthur B. McDonald.
Hypothesis	Arthur B. McDonald is a Japanese physicist, known for neutrino experiments at the Kamiokande and its successor, Super-Kamiokande.
Label	2 (Contradiction)
Predicted Label	0 (Entailment)

Table 3: An example of the most common type of mistake made by the model: Contradiction but Predicted Entailment.

Is our baseline model learning reasoning or is it simply making predictions based on something like token overlap? Taking a look at all errors made by the model, we can see that on average, the word overlap between premise and hypothesis is much higher where

the model predicts entailment. In comparison to the ground truth labels, the highest average overlap is actually in the contradictions.

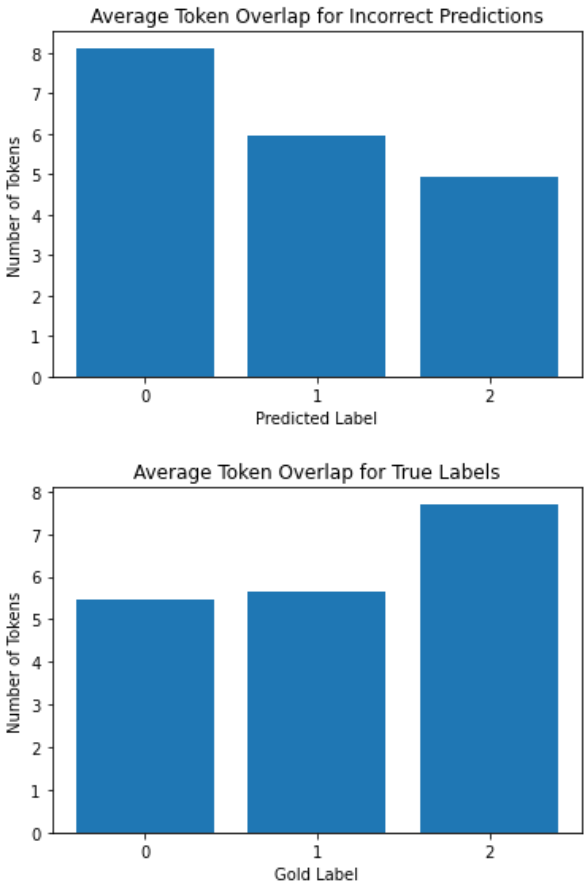

Figure 2: We can see that in the top figure there is a higher average token overlap when the model is predicting entailment and is incorrect, thus implying that the model is more likely to incorrectly predict entailment when it sees a higher overlap of words. The bottom figure demonstrates that in reality contradictions on average have higher word overlap between premise and hypothesis.

Looking at the example given in Table 3, the overlap between premise and hypothesis is essentially the entire hypothesis, which could be the reason why the model is incorrectly labeling it as an entailment instead of a contradiction.

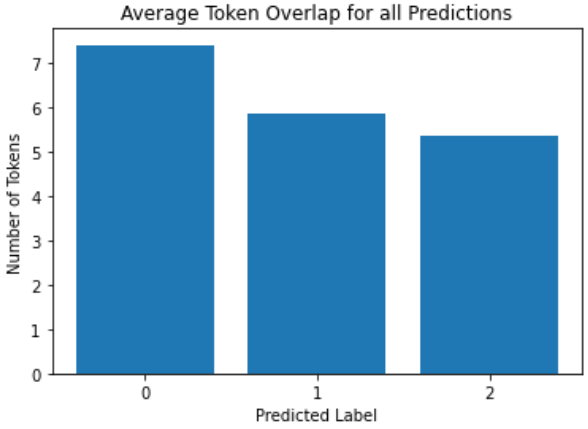

Figure 3: Above displays average token overlap between premise and hypothesis pairs for all predictions made by the model, demonstrating that there is some bias in predicting entailment when the hypothesis is syntactically similar to the premise.

2.2 Model Improvements

With an accuracy of less than 30% on the adversarial evaluation set, there is a lot of room for improvement as simply guessing entailment for every example would produce a better accuracy score. The method that we chose to utilize to improve the model performance on adversarial data is by fine-tune training on adversarial data in the same way presented in Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets (Liu et al., NAACL 2019). The fine-tuning was done using one of the training sets from the ANLI dataset (Nie et al., ACL 2020), containing roughly 45,000 example pairs. We fine-tuned using 3 epochs. After fine-tuning was complete we saw that while the overall accuracy increased, the new model does not perform significantly better on contradictions, which is the error type we

were doing most of our analysis on. It failed to correct the example provided in Table 3 as the model still deemed it an entailment instead of a contradiction. Here is another example of where the model failed in both situations:

Premise: Suzy Nakamura is an American actress. Nakamura is known for her many guest appearances on sitcoms such as "According to Jim", "Half and Half", "8 Simple Rules", "Curb Your Enthusiasm" and "How I Met Your Mother" and her recurring role in the early seasons of the drama "The West Wing" as assistant to the Sam Seaborn character, as well as Dr. Miura in the ABC sitcom "Modern Family".'

Hypothesis: Suzy Nakamura is an American actress who had a recurring role as Dr. Miura in the early seasons of the drama "The West Wing.

Gold Label: 2

Predicted Label: 0

To the human eye, the hypothesis is a contradiction to the premise as Suzy Nakamura did not play Dr. Miura in “West Wing,” but rather in “Modern Family.” This is another case of us wondering if the model is simply learning the frequency of word overlaps between premise-hypothesis pairs as the hypothesis is very similar to words seen in the premise.

Nonetheless, overall performance did increase significantly across the board, especially for the entailment and neutral classes. Overall accuracy increased from **29.9% to 49.9%**.

	Baseline	**Fine-Tuned**
Entailment	32.3%	60.5%
Neutral	34.2%	55.0%
Contradiction	23.1%	34.2%

Table 4: Accuracy scores for each label group before and after fine-tuning the model on adversarial examples.

Though the model was not effective in addressing the errors we were originally targeting, it was able to correct a large number of entailments. Here is an example of such correction:

Premise: Magnus is a Belgian joint dance project of Tom Barman (from the rock band dEUS) and CJ Bolland. Magnus\' debut album, "The Body Gave You Everything", was released on March 29, 2004. Two of its tracks, "Summer\'s Here" and "Jumpneedle", were released as singles.

Hypothesis: The body gave you everything" album was not released on March 28, 2003 but on March 29, 2004.

With the baseline model, this pair was predicted to be a contradiction, but with the fine-tuned model, it was correctly classified as an entailment.

We then used the fine-tuned model to do inference on the original, non-adversarial evaluation set and found that our overall accuracy decreased from **89.3% to 82.8%**. This now raises the question: Is a 20% increase in accuracy on an infrequent task worth losing nearly 7% on a simpler, more frequent task? The simple answer is that it depends on what the model is being used for.

2.3 Replication on Sarcasm

After finishing analysis on the ANLI dataset (Nie et al., ACL 2020), we wanted to see if we could replicate the results on a more specific adversarial task. The first challenge that came to mind was sarcasm due to the fact that by nature, sarcastic sentences mean the opposite of what is being said. Sarcasm often confuses many humans, let alone a deep learning model. We began by generating an evaluation set of 165 pairs whose premise contained sarcasm. Here is an example:

Premise: This meeting could not be more engaging. I've only checked the clock five times in the last minute.

Hypothesis: You found the meeting extremely captivating.

Label: 2

This is clearly a contradiction as the sarcasm in the premise indicates that the person could not wait for the meeting to end.

We did inference using our baseline model on this sarcasm adversarial set and saw that the model performed poorly with an overall accuracy of **29.7%**. The model predicted the example used above as an entailment. We then created a training set using similar examples with sarcasm in the premise. In total we generated 400 pairs and used this set to fine-tune our baseline model. With our new model, we did inference on the same evaluation set and saw that accuracy increased to **76.4%** while only decreasing accuracy on the original evaluation set from **89.3% to 88.3%**.

	Baseline	Fine-Tuned
Entailment	15.0%	96.3%
Neutral	72.2%	47.2%
Contradiction	22.4%	65.3%

Table 5: Accuracy scores for each label group before and after fine-tuning the model on sarcastic examples.

3 Results

As demonstrated in the previous section, it has been proven effective to fine-tune models on adversarial data in order to improve performance on a similar evaluation set.

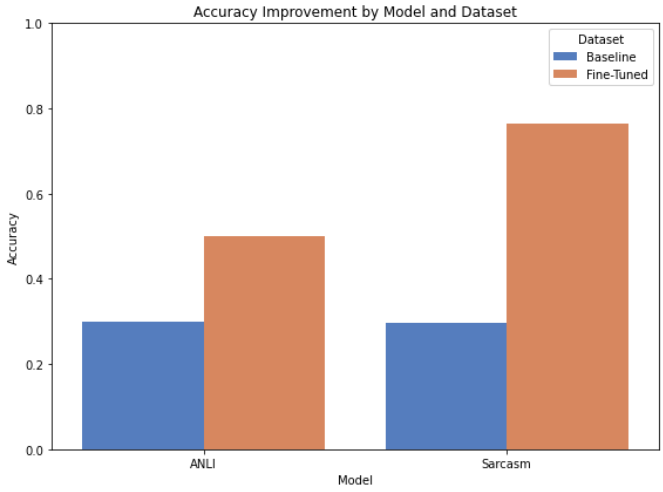

Figure 4: Figure above displays the improvement in accuracy on each evaluation set after fine-tuning each of the models.

3.1 Analysis of Results

While both fine-tuned models were able to achieve higher accuracy, we believe that the accuracy on the sarcasm evaluation set had a much larger increase due to the fact that the generated sarcasm data has less variation in it

than ANLI which contains a whole host of challenging premise-hypothesis pairs. The sarcasm data was more simple in that it contains a sarcastic remark for the premise and the hypothesis interprets whether the individual is enjoying the activity or not.

As for what these models are truly learning, we are confident in saying that fine-tuning the models allowed us to remove bias that may have been caused by the word overlap between premises and hypotheses. To reiterate, Figure 3 demonstrates that when the baseline model predicts entailment, the overlap between the premise and hypothesis is on average higher. This intuitively makes sense as the more similar sentences are, the more likely they are to have the same meaning. After fine-tuning the baseline model on adversarial data, we saw overlap averages closer to the averages of the true labels.

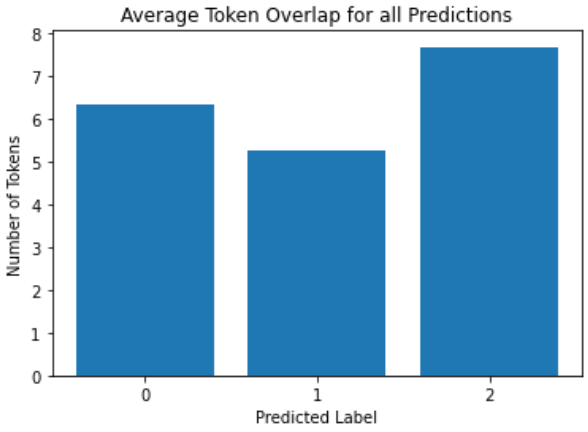

Figure 5: Predicted label token overlap averages between premise and hypothesis after fine-tuning the model on adversarial data. Similar to true averages as displayed in the bottom graph of Figure 2.

The same bias was corrected when fine-tuning the model on sarcasm data. When doing inference on the sarcasm evaluation set using the baseline model, average overlap between premise and hypothesis was highest amongst the entailments. After fine-tuning the model on sarcasm data, the overlap averages for predicted labels were also closer to that of the true label averages.

4 Conclusion

We determined that state-of-the-art NLP models such as Electra perform poorly when doing inference on adversarial evaluation sets. Through our experiments we also discovered that when performing natural language inference tasks, Electra is inclined to predict entailment when there is a major overlap in words between the premise and hypothesis. We discovered a way to mitigate this bias and improve accuracy. Both of our models saw significant improvements in accuracy after fine-tuning each model on their respective adversarial challenge sets. The model fine-tuned on the ANLI dataset increased accuracy on the evaluation set by 20% while the model tuned on the sarcasm set saw an increase in accuracy of nearly 47%. Even after fine-tuning these models with challenging data, performance on the original evaluation set remained strong. While this is a good first step, more work needs to be done to understand what these models are truly learning, which will help us build better models that are capable of comprehending all the nuances of the human language.

References

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).

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

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. Association for Computational Linguistics.