Reinforcing Noise Robustness in Natural Language Inference Tasks

Abstract

This work explores the robustness of a Natural Language Inference (NLI) model, based on the ELECTRA architecture, when exposed to varying noise conditions. We methodically assess the performance of the model under character-level, word-level, and combined noise, applied to the premise, hypothesis, and both sentences, also across a range of perturbation probabilities. Findings show notable sensitivity to noise, especially combined noise and noise introduced in the hypothesis. To address these vulnerabilities, inoculation is utilized, by introducing controlled noise into the training data, it allows the model to better generalize under noisy conditions, while simultaneously preserving general accuracy on clean data. After training, results demonstrate that inoculation effectively enhanced the model's toughness against noise, as well as specific types of it.

1 Introduction

Pre-trained language models, such as BERT, and their variants, have become fundamental in Natural Language Processing (NLP) due to their impressive performance on a wide range of benchmark datasets. However, despite their effectiveness on clean and curated datasets, these models often struggle when dealing with data from the real-world, which can be noisy, inconsistent, and/or otherwise perturbed. This lack of robustness is particularly concerning in tasks such as Natural Language Inference (NLI), where changes to the input can impact the performance of the model.

In this paper, we evaluate the robustness of the ELECTRA small model (Clark et. Al., 2020) trained for NLI with The Stanford NLI dataset (Bowman et al., 2015). Several analysis methods are conducted on the model with different types of noise, on different locations, across a range of noise probabilities. This approach allows us to analyze the model's sensitivity to specific noise types and application locations, helping us uncover artifacts in the model's behavior that contribute to performance degradation.

To address the identified robustness issues, we employ an inoculation strategy (Liu et al., 2019). This approach involves gradually incorporating controlled noise into the training data, with the goal of enhancing the model's resilience to noisy inputs. By exposing the model to noise during training, we aim to improve its generalization.

2 Base Model Analysis

2.1 Experiment Setup

To evaluate the robustness of the base ELECTRA model for Natural Language Inference (NLI). The base model achieves an ~89% accuracy on non-perturbed datasets. We systematically introduced three types of noise: character-level, word-level, and combined noise. Noise was applied to either the premise, hypothesis, or both components of the NLI task at varying probabilities (from 0% to 100%). We measured the model's performance in terms of accuracy and loss to observe its sensitivity to these noise variations.

2.2 Analysis Methods

To reveal any artifacts in the model's behavior, we utilize methods inspired by contrast sets (Gardner et al., 2020) and adversarial challenge sets (Jia and Liang, 2017; Wallace et al., 2019).

Contrast Sets: By applying specific types of noise (character, word, and combined) to different sentence components (premise, hypothesis, or both), we created a set of controlled contrasting conditions. Specifically, there is a probability that either a character or word is added, deleted, or switched in the sequence. This allowed us to observe how minor variations in the input affected the model's predictions, revealing vulnerabilities specific to each noise type and application location.

Adversarial Challenge Sets: Additionally, we introduced challenging noise patterns by combining multiple noise types and applying them to the premise, hypothesis, and both. This approach, akin to adversarial testing, aimed to stress-test the model's robustness under compounded disturbances, highlighting significant weaknesses in its handling of noisy data.

2.3 Impact of Noise Type and Location

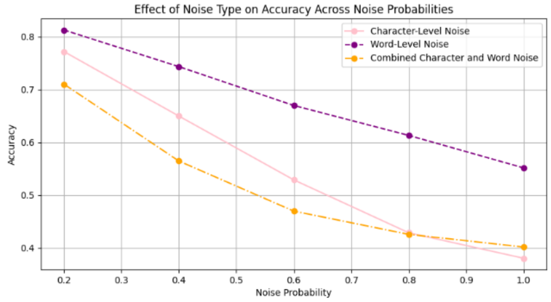

Figure 1. A graph showing the effect of character-level, word-level, and combined noise on accuracy across noise probabilities.

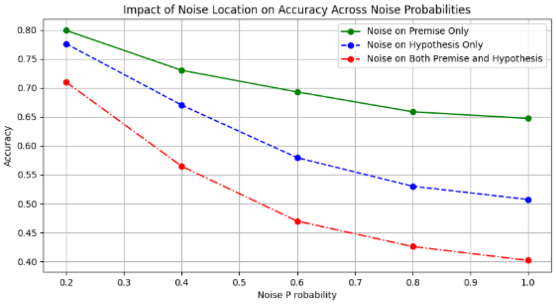

Figure 2. Graph showing the impact of noise location on accuracy across noise probabilities: noise applied to premise only, hypothesis only, and both.

Figure 1 and Figure 2 illustrate the model's sensitivity to different noise types and locations, respectively. Word-level and combined noise degraded performance more rapidly than character-level noise, suggesting that disruptions at the word level have a stronger effect on sentence semantics. Additionally, the model exhibited increased vulnerability when noise was applied to the hypothesis, likely due to its role in inference.

When noise affected both the premise and hypothesis, performance declined most severely, highlighting the model's compounded sensitivity to disturbances across both components. As such we train the models for the worst-case scenarios.

2.4 Performance Overview

Our analysis revealed that as noise probability increased, both accuracy and loss degraded significantly. Figure 3 below shows the model's overall performance under combined noise applied to both the premise and hypothesis, illustrating a steady decline in accuracy as noise levels rise. The model showed some resilience to low-level character noise, but word-level and combined noise resulted in more pronounced performance drops.

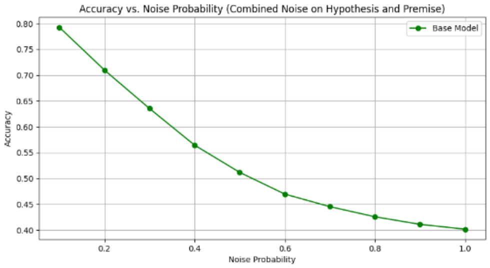

Figure 3. Graph showing model accuracy across noise probabilities with combined noise applied to both the hypothesis and premise.

2.5 Discussion

These results show critical artifacts in the model's performance. Character-level noise often led to vocabulary related errors, where changed tokens were misinterpreted. Word-level perturbations caused semantic confusion, particularly under combined noise on both premise and hypothesis sentences, suggesting a breakdown in sentence clarity. These findings underscore the model's general inflexibility to noise, especially at the word level and in cases where both the premise and hypothesis are affected. Our inoculation approach aims to address these challenges.

3 Inoculation

Given the base model’s sensitivity to noise, particularly a combined noise affecting both the premise and hypothesis sentences, we implemented an inoculation strategy to improve robustness. This strategy involves training the model on controlled noisy data to expose it to the types of perturbations it may encounter.

We introduced controlled noise into the training data, with each noise type (character-level, word-level, and combined) applied in different arrangements (premise only, hypothesis only, both). Noise probabilities were varied, with models trained on 20% and 40% noisy data to assess the impact of different inoculation levels. Each inoculated model was then evaluated under the same noise conditions used to test the base model. Performance was measured in terms of accuracy and loss across different noise probabilities and noise types, allowing us to quantify improvements in robustness and identify the optimal inoculation level. We hypothesized that inoculation would improve resilience to noisy inputs without compromising accuracy on clean data.

4. Results

We examined the performance of the inoculated models (trained on 20% and 40% noisy data) against the base model under various noise conditions to assess improvements in robustness. Our results show that inoculation does in fact increase the model's performance across noise levels, with inoculated models retaining higher accuracy compared to the base model, especially under combined noise applied to both the premise and hypothesis sentences.

4.1. Performance Across Noise Levels

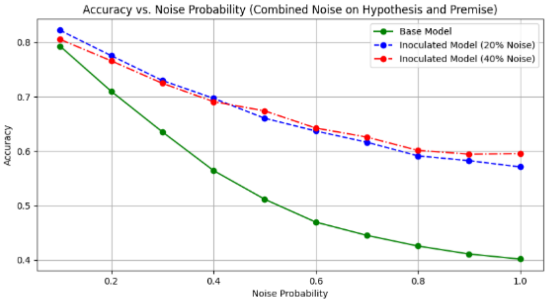

Figure 4: Graph showing the accuracy trends for the base model and the two inoculated models across noise probabilities under combined noise applied to both the premise and hypothesis. The base model's accuracy drops sharply as noise probability increases, while the inoculated models, particularly the one trained with 40% noise, demonstrate more gradual degradation.

This comparison indicates that training with noisy data significantly enhances robustness to perturbations, especially at higher frequencies. In low noise environments, the 20% inoculated model performed slightly better than the two other models. In the mid-range noise probabilities, the 40% noise inoculated model and the 20% inoculated model both have similar accuracies. And in the higher range, the 40% model slightly outperforms the 20%, suggesting that a higher level of inoculation further strengthens resilience to noise.

4.2. Analysis of Misclassification Patterns

To gain a deeper understanding of how inoculation impacts specific error types, we examined the misclassification rates across the three models under varying noise levels. Figure 5 below presents heatmaps showing the distribution of misclassifications for each model across different label transitions (e.g., Entailment → Neutral, Neutral → Contradiction).

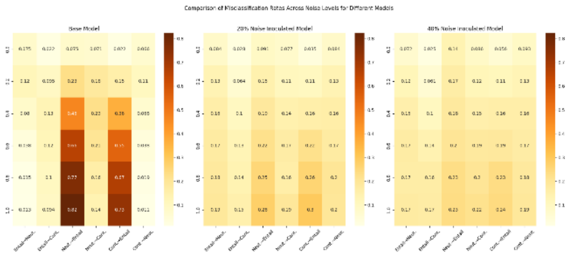

Figure 5: Several heatmaps comparing the misclassification rates across noise levels for base and inoculated models.

Character-Level Analysis: For the base model, noise frequently leads to misclassifications where entailment labels are incorrectly classified as neutral or contradiction. The inoculated models display lower misclassification rates in these areas, particularly with 40% inoculation, indicating that inoculation reduces the model's tendency to make these specific errors under noisy conditions.

Word-Level and Combined Noise Analysis: The heatmaps reveal that the base model struggles particularly with combined noise, leading to frequent misclassification across all categories. However, the 40% noise inoculated model shows a noticeable reduction in these errors, especially for entailment and contradiction pairs, suggesting that inoculation is effective at reducing noise-induced misinterpretations.

4.3 Key Observations

Both inoculated models, especially the one trained on 40% noise, show improved robustness across the board as well as a significantly lower misclassification rate compared to the base model under perturbed conditions.

Inoculation reduces the model's sensitivity to the combined type of noise, which was, as shown previously, the most debilitating noise condition for the base model. The inoculated models maintain higher accuracy and make fewer severe misclassifications under these conditions.

The method also seems to help the model keep relationships between entailment, neutral, and contradiction even under perturbations. This suggests that this type of training helps the model to better manage disruptions in input, preserving the semantic relationships essential for NLI tasks.

5 Conclusion

This work explored the efficacy and robustness of an ELECTRA-based model tasked with Natural Language Inference (NLI) under different levels and types of noise on different areas of the sequences. Initial analysis revealed that the model performed poorly to word-level noise and noise combined with characters, especially when it affects both premise and hypothesis sentences at high noise probabilities. To address these issues exposed by our investigation, we utilize a method called inoculation, where different models are trained with controlled levels of noise (20% and 40%) in the training data. The following results showed that the new models performed more superiorly than the base model under noise across all probabilities. The inoculated models demonstrated their robustness along with their ability to retain semantic understanding by showing reduced misclassification rates even under perturbations. Real-world applications of inoculation in models have real potential as, unlike in simulations, noise is everywhere. In the future, we could explore more nuanced inoculation strategies, and test and introduce more intuitive types of noises

References

[Bowman et al.2015] 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, September. Association for Computational Linguistics.

[Clark et al.2020] 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)

[Gardner et al.2020] 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.

[Jia and Liang 2017] 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.

[Wallace et al.2019] 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.