Interpreting the Robustness of Neural NLP Models to Textual Perturbations

Yunxiang Zhang[1], Liangming Pan[2], Samson Tan[2], Min-Yen Kan[2]
[1]Peking University, [2]National University of Singapore

Abstract

Modern Natural Language Processing (NLP) models are known to be sensitive to input perturbations and their performance can decrease when applied to real-world, noisy data. However, it is still unclear why models are less robust to some perturbations than others. In this work, we test the hypothesis that the extent to which a model is affected by an unseen textual perturbation (robustness) can be explained by the learnability of the perturbation (defined as how well the model learns to identify the perturbation with a small amount of evidence). We further give a causal justification for the learnability metric. We conduct extensive experiments with four prominent NLP models -- TextRNN, BERT, RoBERTa and XLNet -- over eight types of textual perturbations on three datasets. We show that a model which is better at identifying a perturbation (higher learnability) becomes worse at ignoring such a perturbation at test time (lower robustness), providing empirical support for our hypothesis.

Introduction

A robust NLP model should not be easily fooled by slight noise in the text. Given the difference of robustness between models and perturbations, it is a natural question why models are more sensitive to some perturbations than others. It is crucial to avoid over-sensitivity to input perturbations, and understanding why it happens is useful for revealing the weaknesses of current models and designing more robust training methods. To the best of our knowledge, a quantitative measure to interpret the robustness of NLP models to textual perturbations has yet to be proposed. To improve the robustness under perturbation, it is common practice to leverage data augmentation. Similarly, how much data augmentation through the perturbation improves model robustness varies between models and perturbations. In this work, we aim to investigate two **Research Questions (RQ)**:

- **RQ1:** *Why are NLP models less robust to some perturbations than others?*
- **RQ2:** *Why does data augmentation work better at improving the model robustness to some perturbations than others?*

Setup and Terminology

Setup: As a pilot study, we consider the task of binary text classification.
Perturbation: A transformation that injects a specific type of noise into a piece of text (Figure 1).
Robustness: We apply the perturbations to the test set and measure the robustness of the model to a perturbation as the decrease in accuracy.
Post Augmentation Δ: We simulate the data augmentation process by appending perturbed data to the training set. We calculate the improvement in performance after data augmentation as the difference of test accuracies.

Learnability Hypothesis

Learnability: We want to compare perturbations in terms of how well the model learns to identify them with a small amount of evidence. We cast learnability estimation as a perturbation classification task, where a model is trained to identify the perturbation in an example. We define that the learnability estimation consists of three steps:

1. **Assigning random labels.** We randomly assign pseudo labels to each training example regardless of its original label.
2. **Perturbing with probabilities.** We apply the perturbation to each training example in one of the pseudo groups.
3. **Estimating model performance.** We train a model on the randomly labeled dataset with perturbed examples. The perturbation learnability is the difference of accuracies on perturbed and unperturbed test set with random pseudo labels.

We propose hypotheses for RQ1 and RQ2:

- **Hypothesis 1 (H1):** *A model for which a perturbation is more learnable is less robust against the same perturbation at the test time.*
- **Hypothesis 2 (H2):** *A model for which a perturbation is more learnable experiences bigger robustness gains with data augmentation along such a perturbation.*

Causal Explanation

Motivation. Learnability is motivated by concepts from the causality literature. In fact, learnability is the causal effect of perturbation (treatment) on model predictions (outcome), which is often difficult to measure due to the confounding latent features.
A Causal Explanation for Random Label Assignment. Why do we assign random labels before perturbations? Because randomization decouples the effects of perturbation and other confounding latent features (Figure 2). As a result, we can directly calculate the causal effect from the observed outcome, which is exactly the difference of model accuracy on the perturbed and unperturbed test sets with random labels.
Learnability is a Causal Estimand. We further identify learnability as a causal estimand, Average Treatment Effect (ATE), which is a measure used to compare treatments in randomized experiments.

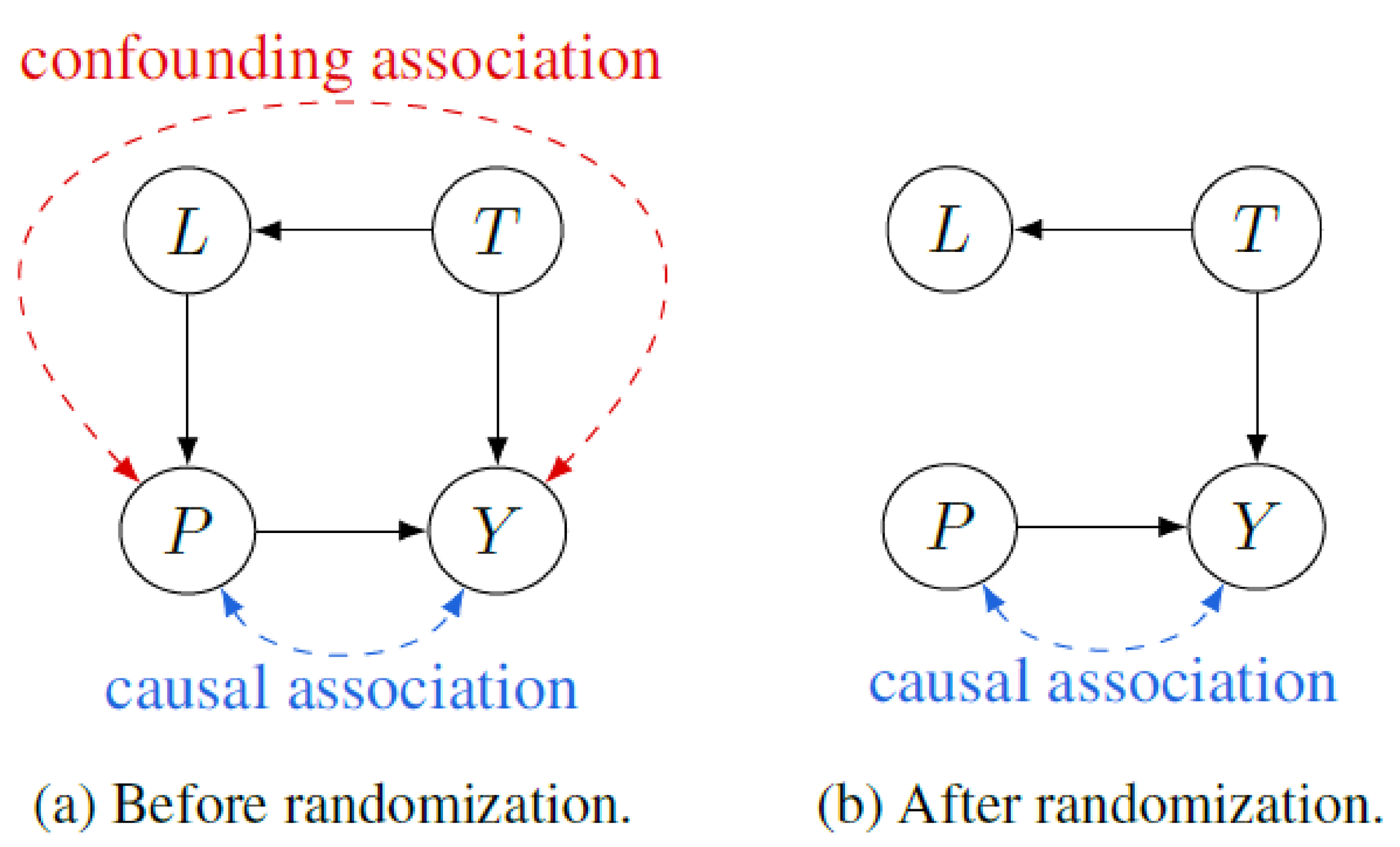

Figure 2. A causal graph explanation for decoupling perturbation and latent feature with randomization. P is the perturbation and T is the latent feature. L is the original label and Y is the predicted label.

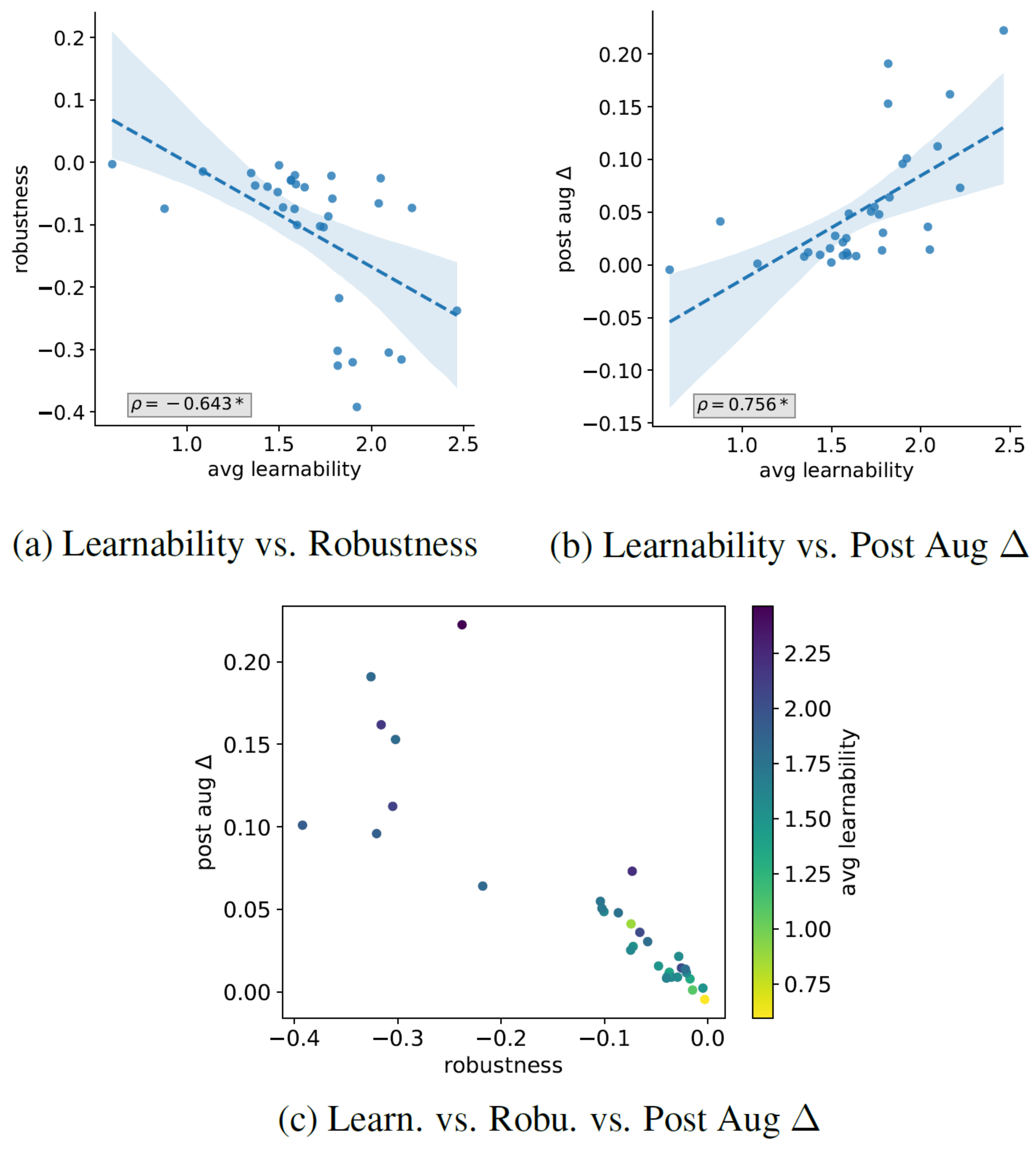

Figure 3. Linear regression plots of learnability vs. robustness vs. post data augmentation Δ on IMDB dataset. Each point in the plots represents a model-perturbation pair. ρ is Spearman correlation. * indicates high significance (p-value < 0.001).

Experiment & Result

Experimental Settings. To test the learnability, robustness and improvement by data augmentation with different NLP models and perturbations, we experiment with four modern and representative neural NLP models: TextRNN, BERT, RoBERTa and XLNet. We use three common binary text classification datasets --- IMDB movie reviews (IMDB, Yelp polarity reviews (YELP), Quora Question Pair (QQP) --- as our testbeds. We select eight character-level or word-level perturbation methods in existing literature (Figure 1) that simulate different types of noise an NLP model may encounter in real-world situations.
Empirical Findings. For RQ1, we observe a negative correlation between learnability and robustness (Figure 3a), validating Hypothesis 1. For RQ2, we find that data augmentation with a perturbation the model is less robust to has more improvement on robustness (Figure 3b), validating Hypothesis 2. Combining these two findings (Figure 3c), we further show that data augmentation is *only* more effective at improving robustness against perturbations that a model is more sensitive to.

Conclusion

This work provides an empirical explanation for why NLP models are less robust to some perturbations than others. The key to this question is perturbation learnability, which is grounded in the causality framework. We find that learnability, which causally quantifies how well a model learns to identify a perturbation, is predictive of the model robustness to the perturbation. We also show that data augmentation is only more effective at improving robustness against perturbations that a model is more sensitive to.

Perturbation	Example Sentence
None	His quiet and straightforward demeanor was rare then and would be today.
duplicate_punctuations	His quiet and straightforward demeanor was rare then and would be today..
butter_fingers_perturbation	His quiet and straightforward demeanor was rarw then and would be today.
shuffle_word	quiet would and was be and straightforward then demeanor His today. rare
random_upper_transformation	His quiEt and straightForwARd Demeanor was rare TheN and would be today.
insert_abbreviation	His quiet and straightforward demeanor wuz rare then and would b today.
whitespace_perturbation	His quiet and straightforward demean or wa s rare thenand would be today.
visual_attack_letters	Hiṣ qủiĕt ȁnd straịghṭforwẳrḍ dɛmeanoŕ wâṣ rȁre tḫen and wouḽd þə tɘḓấȳ.
leet_letters	His qui3t and strai9htfor3ard d3m3an0r 3as rar3 t43n and 30uld 63 t0da4.

Figure 1. An example sentence with different types of perturbations.

Contact

Yunxiang Zhang
Peking University
Email: yx.zhang@pku.edu.cn
Website: https://yunx-z.github.io/