Interpreting the Robustness of Neural NLP Models to Textual Perturbations

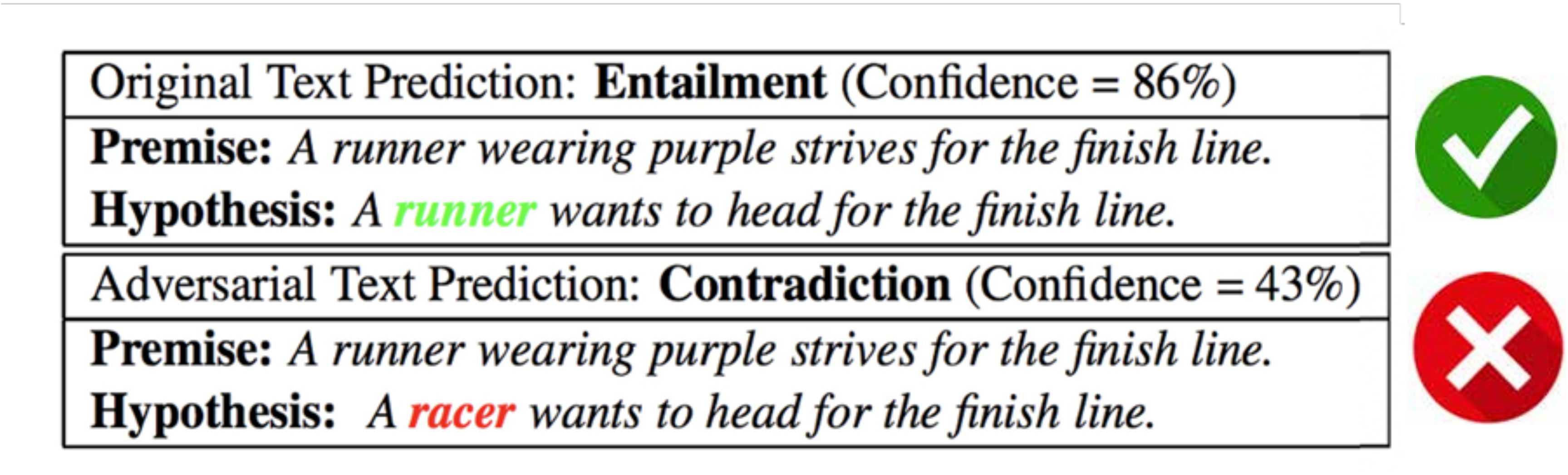
Yunxiang Zhang[1], Liangming Pan[2], Samson Tan[2], Min-Yen Kan[2]

[1]Peking University

[2]National University of Singapore

Findings of ACL 2022

NLP models are less robust to text perturbations



(Prabhakaran et al., 2019; Niu et al., 2020; Ribeiro et al., 2020; Moradi and Samwald, 2021)

Different text perturbation methods

Perturbation	Original text	Perturbed text
Character-level		
Insertion	Who was the first governor of Alaska?	Who was the firsdt governor of Alaska?
Deletion	Mercury, what year was it discovered?	Mercury, what year was it discovred?
Replacement	Who is the Prime Minister of Canada?	Who is the Prime Monister of Canada?
Swapping	What is the primary language in Iceland?	What is the primary lnaguage in Iceland?
Repetition	How many hearts does an octopus have?	How many heartts does an octopus have?
CMW	What kind of gas is in a fluorescent bulb?	What kind of gas is in a florescent bulb?
LCC	How many hearts does an octopus have?	How many hearts does an OCTOPUS have?
Word-level		
Deletion	How much was a ticket for the Titanic?	How much a ticket for the Titanic?
Repetition	What is another name for vitamin B1?	What is another name name for vitamin B1?
RWS	What precious stone is a form of pure carbon?	What valued rock is a form of pure carbon?
Negation	What planet is known as the "red" planet?	What planet is not known as the "red" planet?
SPV	What does a barometer measure?	What do a barometer measure?
Verb tense	Why in tennis are zero points called love?	Why in tennis were zero points called love?
Word order	What is the most common eye color?	What is the common most color eye?

(Milad and Matthias, EMNLP 2021)

Some perturbations are more effective than others

Task	LM	Test set	Character-level perturbation methods						
			Insertion	Deletion	Replace	Swap	Repeat	CMW	LCC
TC	BERT	90.4	77.4	76.2	76.1	76.5	78.8	58.4	78.3
	RoBERTa	**93.1**	79.2	**78.9**	76.3	**76.7**	**80.8**	60.5	78.9
	XLNet	92.0	78.1	78.3	**76.5**	75.2	80.2	61.5	77.4
	ELMo	84.8	**80.4**	78.5	74.7	75.6	79.6	**61.9**	**80.8**

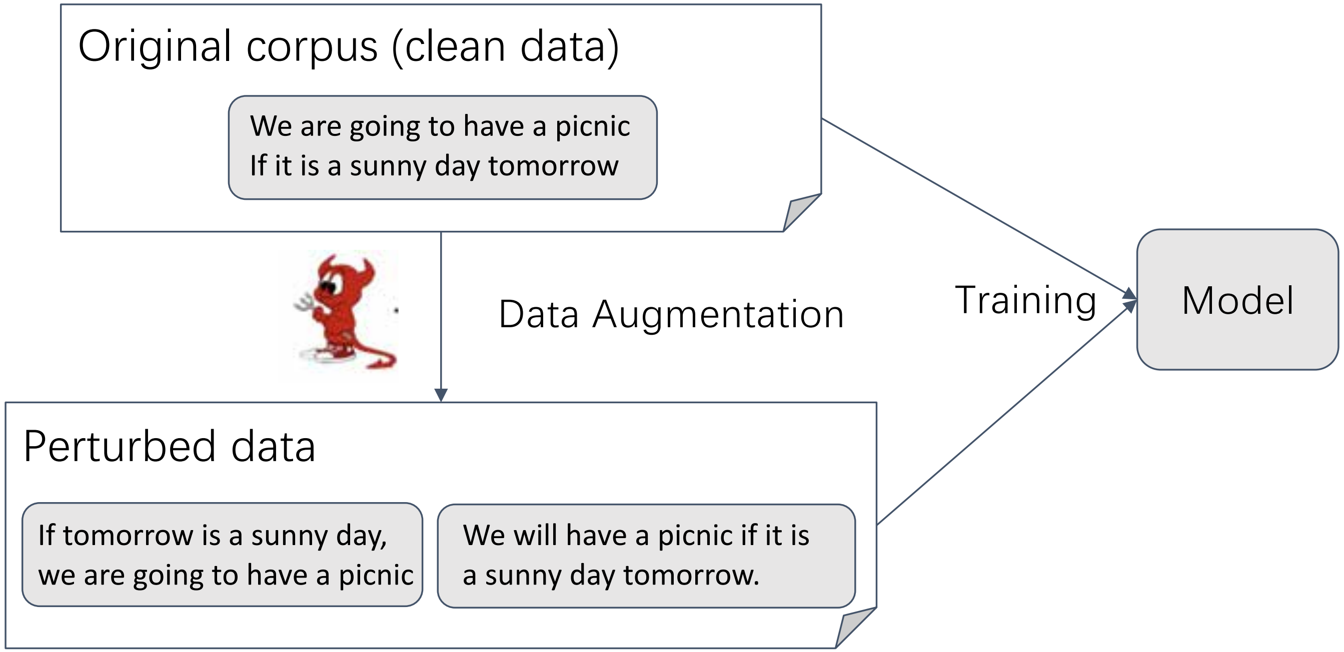
Task	LM	Test set	Word-level perturbation methods						
			Deletion	Repeat	RWS	Negation	SPV	VT	WO
TC	BERT	90.4	75.1	**89.3**	65.7	89.1	88.2	89.0	74.5
	RoBERTa	**93.1**	**76.2**	88.7	73.2	**90.3**	**89.5**	89.4	78.5
	XLNet	92.0	**76.2**	87.5	72.7	89.4	89.0	**89.6**	**83.1**
	ELMo	84.8	72.9	82.8	**75.1**	83.5	83.6	81.2	62.9

Why NLP models are less robust to some perturbations than others?

(Milad and Matthias, EMNLP 2021)

Data augmentation improves the robustness

- To improve the robustness under perturbation, it is common practice to leverage data augmentation.



Data augmentation improves the robustness

- To improve the robustness under perturbation, it is common practice to leverage data augmentation.

- How much data augmentation through the perturbation improves model robustness varies between models and perturbations.

? Why does data augmentation work better at improving the model robustness to some perturbations than others?

Research Questions

RQ1: Why NLP models are less robust to some perturbations than others?

RQ2: Why does data augmentation work better at improving the model robustness to some perturbations than others?

Hypothesis: If the model is more sensitive to a certain kind of perturbation; the model will be less robust to the perturbation. Also, the improvement brought by data augmentation will be more effective.

- Sensitivity is measured by the **Learnability**, which means how well the model can learn to identify the perturbation with a small amount of evidence.

Learnability

The model is more sensitive to a certain kind of perturbation.

↑↓

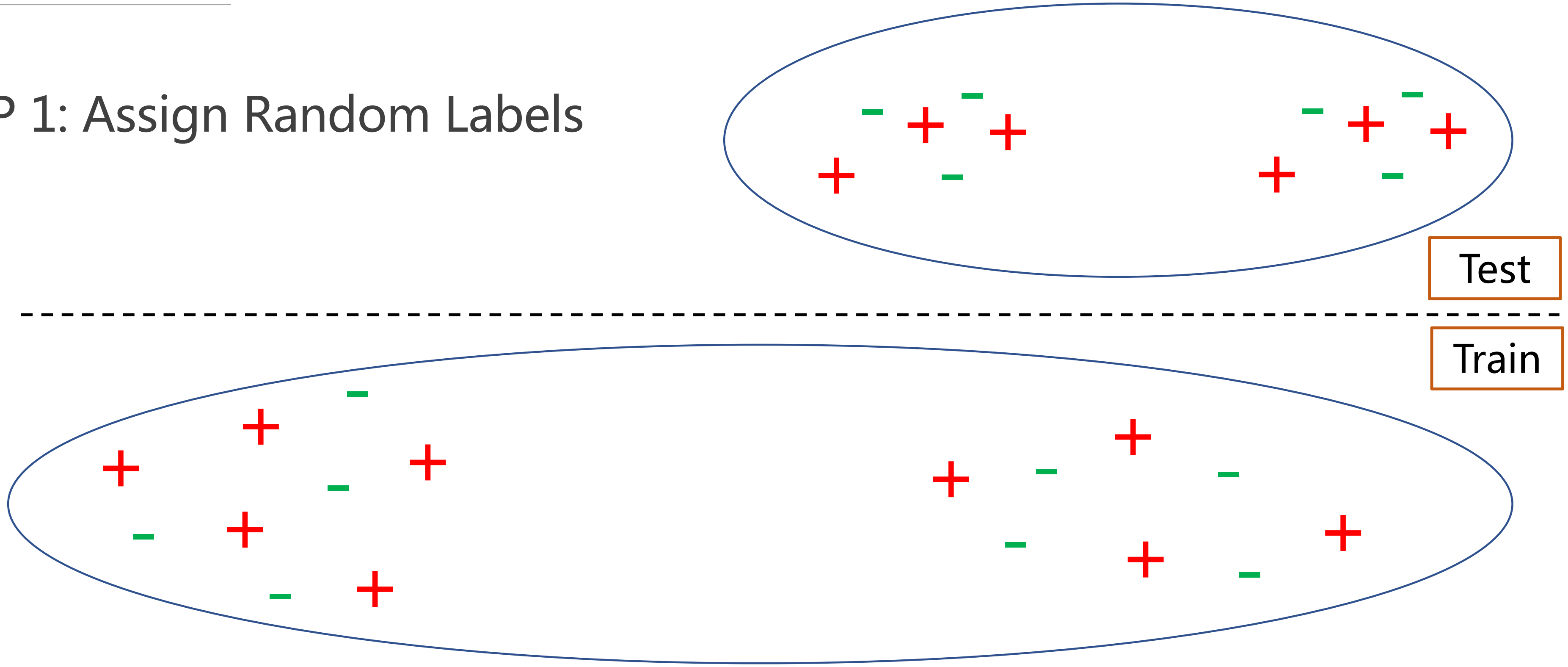
The model is more likely to utilize this spurious feature for prediction.

↑↓

The model can easily learn to identify this perturbation given a small amount of training data.

Learnability Estimation

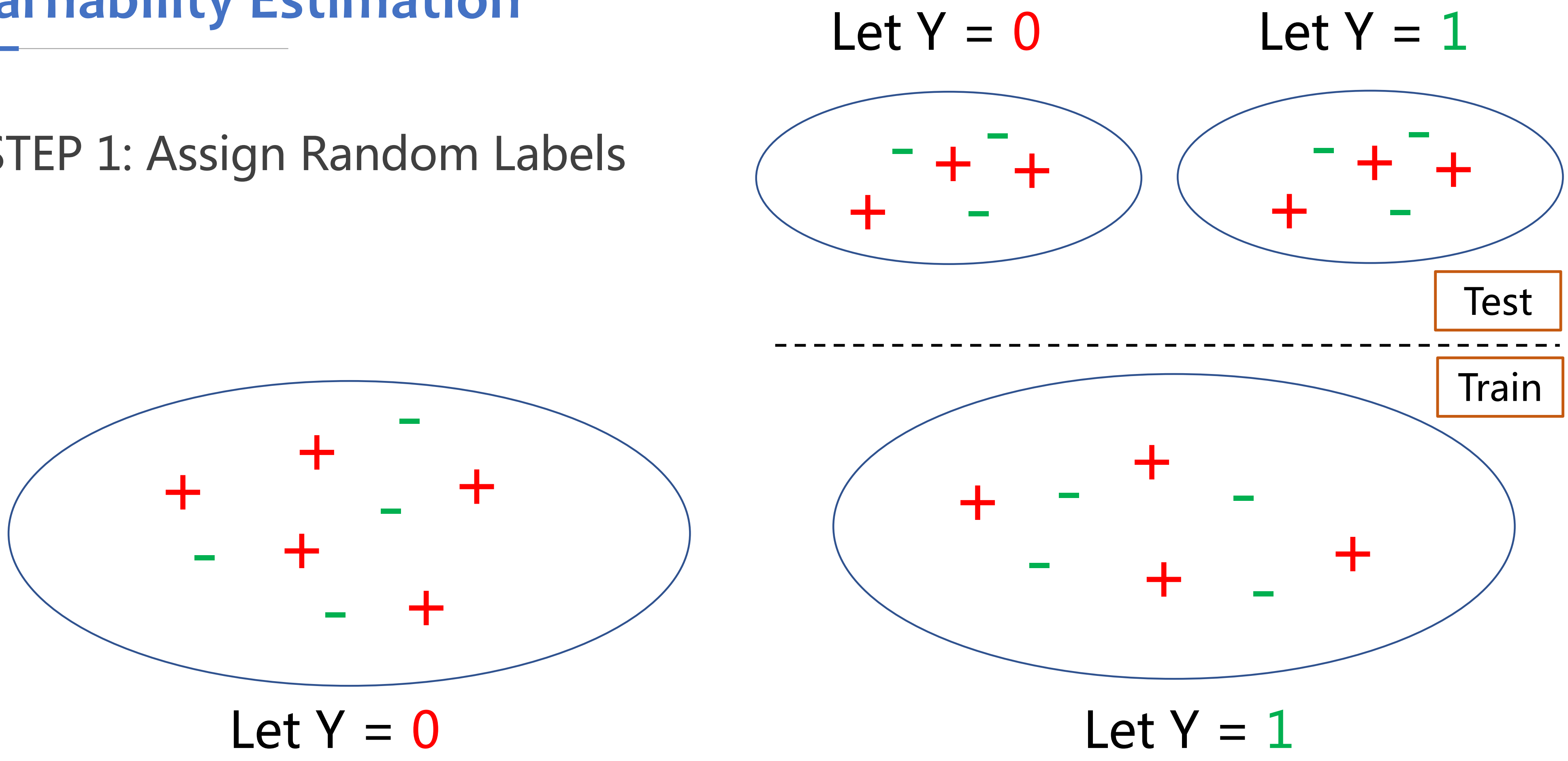
- STEP 1: Assign Random Labels



+ positive example
- negative example

Learnability Estimation

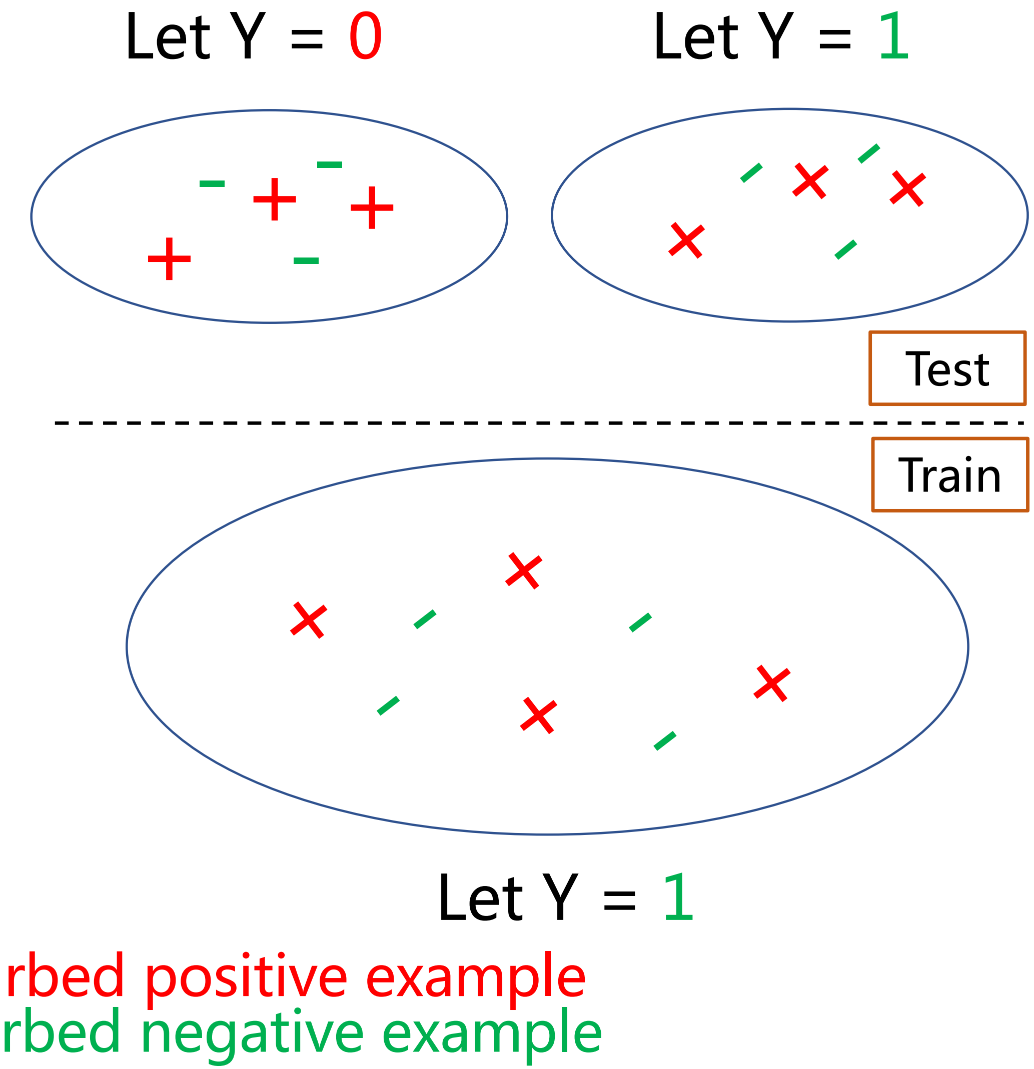
- STEP 1: Assign Random Labels



+ positive example
- negative example

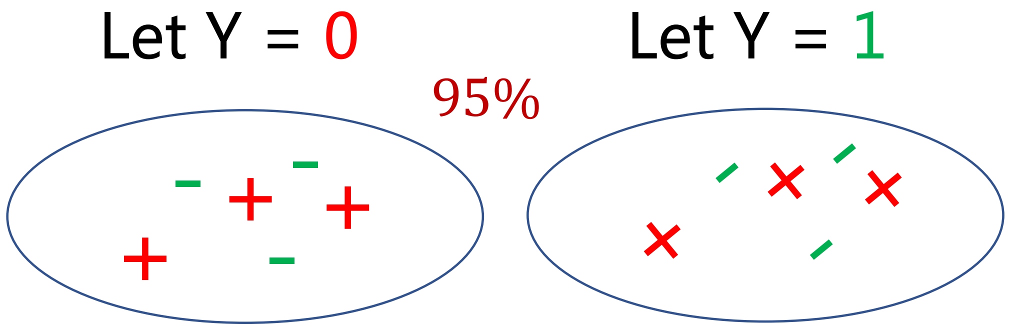
Learnability Estimation

- STEP 1: Assign Random Labels
- STEP 2: Perturb a particular class

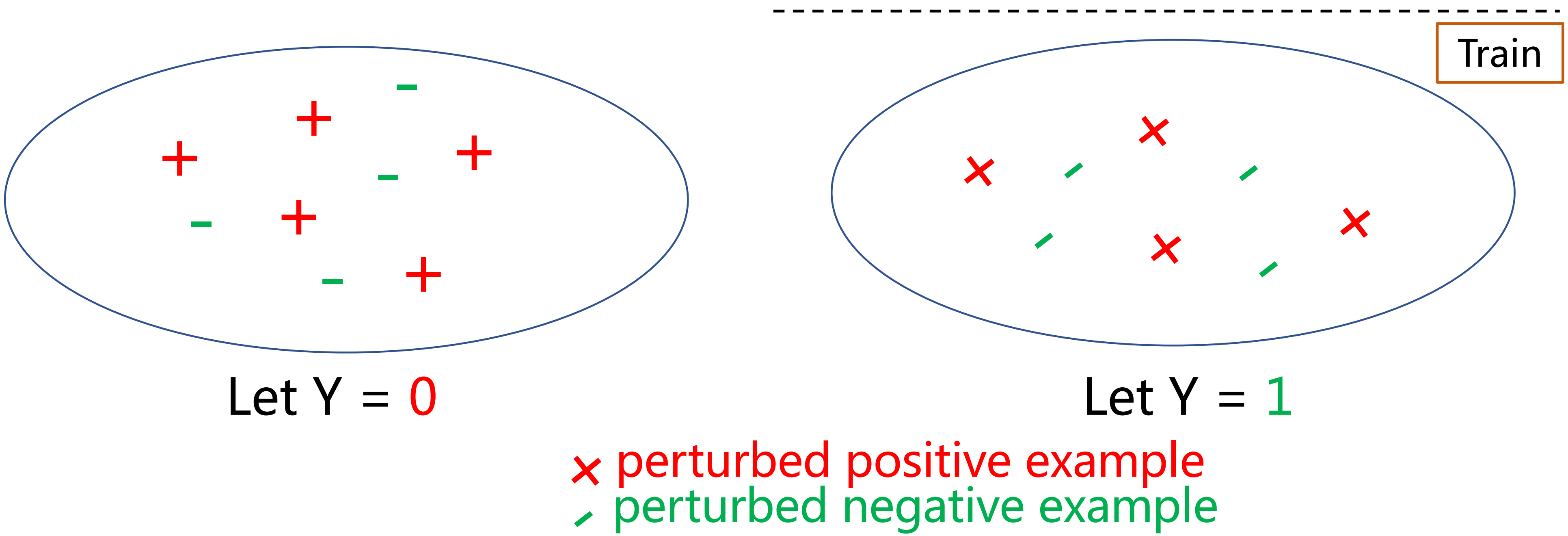


Learnability Estimation

- STEP 1: Assign Random Labels
- STEP 2: Perturb a particular class
- STEP 3: $Learnability$ = accuracy on new test set – original test set



Test



A Causal View

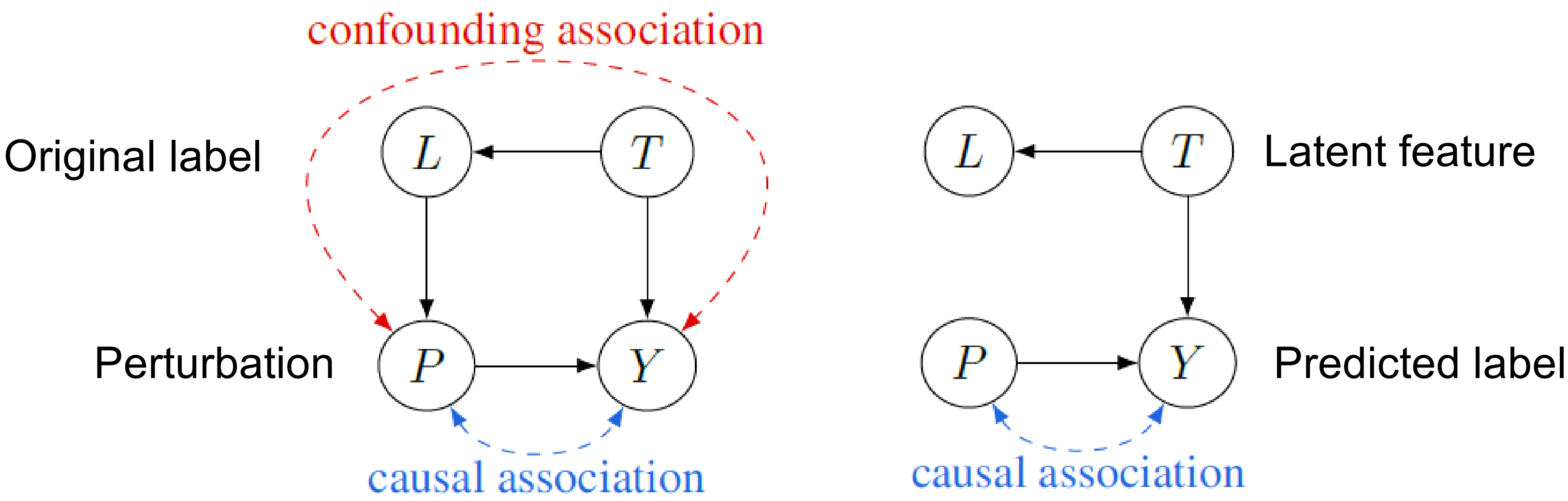
- **Why do we assign random labels before perturbations?**

By randomly assigning pseudo labels to training examples, the only difference between the two pseudo groups is the existence of the perturbation.

- Therefore, the accuracy indicates how well the model can learn to utilize the perturbation for prediction; or in other words, how well the model can learn to identify the perturbed samples.

A Causal View

- **Why do we assign random labels before perturbations?**
  - Randomization decouples the effects of perturbation and other confounding latent features.
  - Learnability is identified as a causal estimand *(Average Treatment Effect, ATE)*



(a) Before randomization. (b) After randomization.

Definitions

Exp No.	Measurement	Label	Perturbation	Training Examples	Test Examples
0	Standard	original	$l \in \varnothing$	$(x_i, 0), (x_j, 1)$	$(x_i, 0), (x_j, 1)$
1	Robustness	original	$l \in \{0, 1\}$	$(x_i, 0), (x_j, 1)$	$(x_i^*, 0), (x_j^*, 1)$

x^* is a perturbed example

$$Robustness = Acc_1 - Acc_0$$

Definitions

Exp No.	Measurement	Label	Perturbation	Training Examples	Test Examples
0	Standard	original	$l \in \varnothing$	$(x_i, 0), (x_j, 1)$	$(x_i, 0), (x_j, 1)$
1	Robustness	original	$l \in \{0, 1\}$	$(x_i, 0), (x_j, 1)$	$(x_i^*, 0), (x_j^*, 1)$
2	Data Augmentation	original	$l \in \{0, 1\}$	$(x_i, 0), (x_j, 1)$ $(x_i^*, 0), (x_j^*, 1)$	$(x_i^*, 0), (x_j^*, 1)$

x^* is a perturbed example

$$\Delta_{post\_aug} = Acc_2 - Acc_1$$

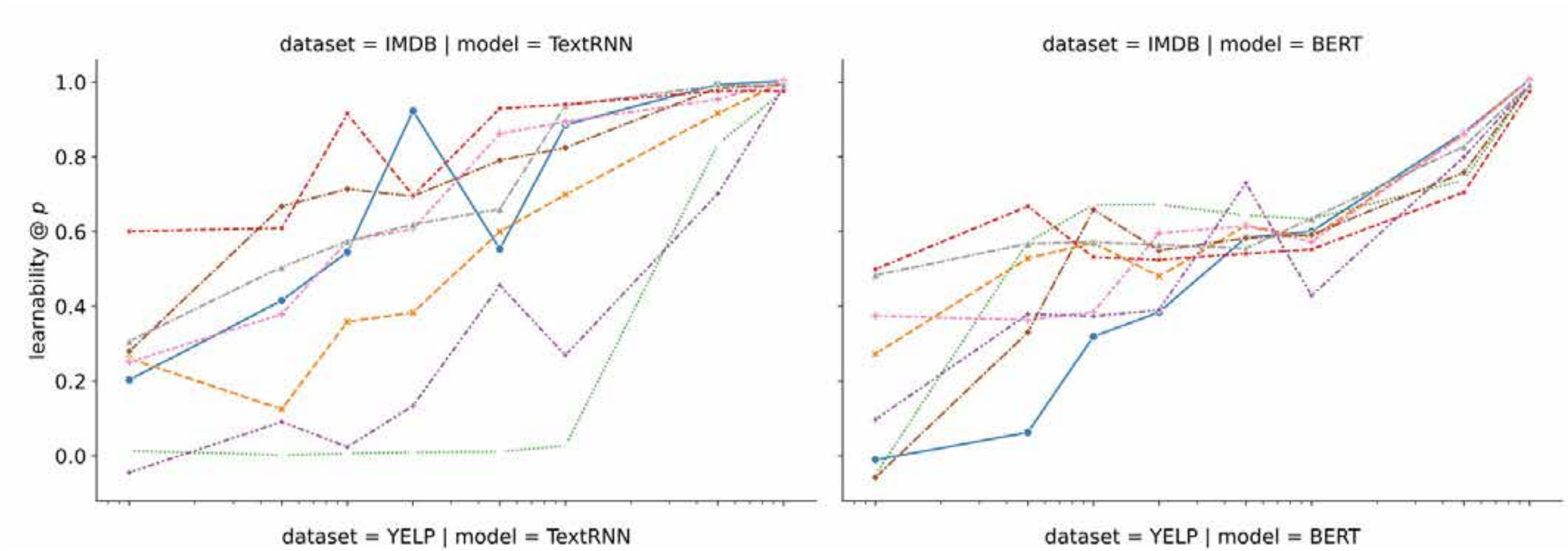
Experiments

- **We estimate robustness, post-augmentation delta, learnability on**
  - Four NLP models: TextRNN, BERT, RoBERTa, XLNet
  - Three datasets: IMDB, YELP, QQP
  - Eight perturbations

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 ẩnd strạighṭḟorwẳrḍ dɛmeanoŕ wâṣ rảre tẖen and wouḷd þə tɘḏẩȳ.
leet_letters	His qui3t and strai9htfor3ard d3m3an0r 3as rar3 t43n and 30uld 63 t0da4.

Results

- Learnability @ p: learnability as a function of perturbation probability.
- We use the AUC (area under curve) to measure the learnability in general.



Results

- Average learnability of each model–perturbation pair on IMDB dataset.
- Different models have different learnability for different perturbations.

Perturbation	XLNet	RoBERTa	BERT	TextRNN	Average over models
whitespace_perturbation	1.638	1.436	1.492	0.878	1.361
shuffle_word	1.740	1.597	1.766	0.594	1.424
duplicate_punctuations	1.086	1.499	1.347	2.050	1.495
butter_fingers_perturbation	1.590	1.369	1.788	1.563	1.578
random_upper_transformation	1.583	1.520	1.721	2.039	1.716
insert_abbreviation	1.783	1.585	1.564	<u>2.219</u>	1.788
visual_attack_letters	**1.824**	<u>1.921</u>	**1.898**	2.094	<u>1.934</u>
leet_letters	<u>1.816</u>	**2.163**	<u>1.817</u>	**2.463**	**2.065**

Results

High learnability: “visual_attack_letters” and “leet_letters”

- They have strong effects on the tokenization process.

Low learnability: “white_space_perturbation” and “duplicate_punctuations”

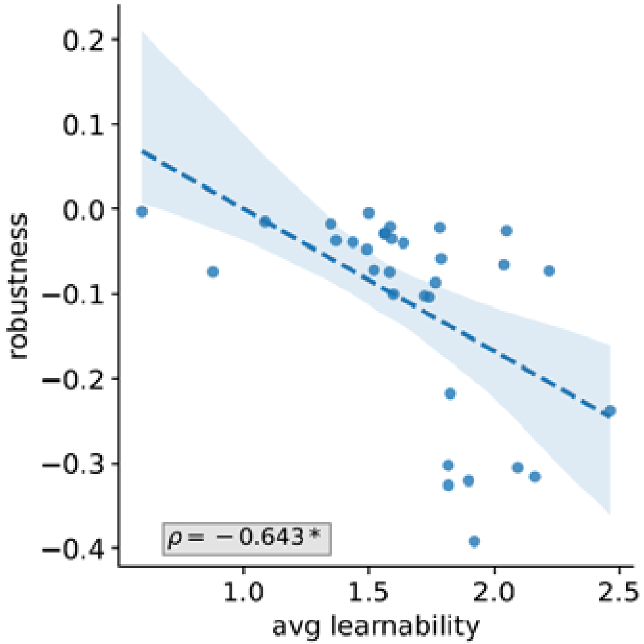
- They have weaker effects on the subword level tokenization, there may already exist similar noise in the pretraining corpora.

Perturbation	XLNet	RoBERTa	BERT	TextRNN	Average over models
whitespace_perturbation	1.638	1.436	1.492	0.878	1.361
shuffle_word	1.740	1.597	1.766	0.594	1.424
duplicate_punctuations	1.086	1.499	1.347	2.050	1.495
butter_fingers_perturbation	1.590	1.369	1.788	1.563	1.578
random_upper_transformation	1.583	1.520	1.721	2.039	1.716
insert_abbreviation	1.783	1.585	1.564	2.219	1.788
visual_attack_letters	**1.824**	1.921	**1.898**	2.094	1.934
leet_letters	1.816	**2.163**	1.817	**2.463**	**2.065**

Results

- We observe a negative correlation between learnability and robustness across all three datasets, validating **Hypothesis 1**.

ρ	IMDB	YELP	QQP
Avg. learnability vs. robustness	$\boxed{-0.643^*}$	-0.821*	-0.695*
Avg. learnability vs. post aug Δ	0.756*	0.846*	0.750*



(a) Learnability vs. Robustness

Results

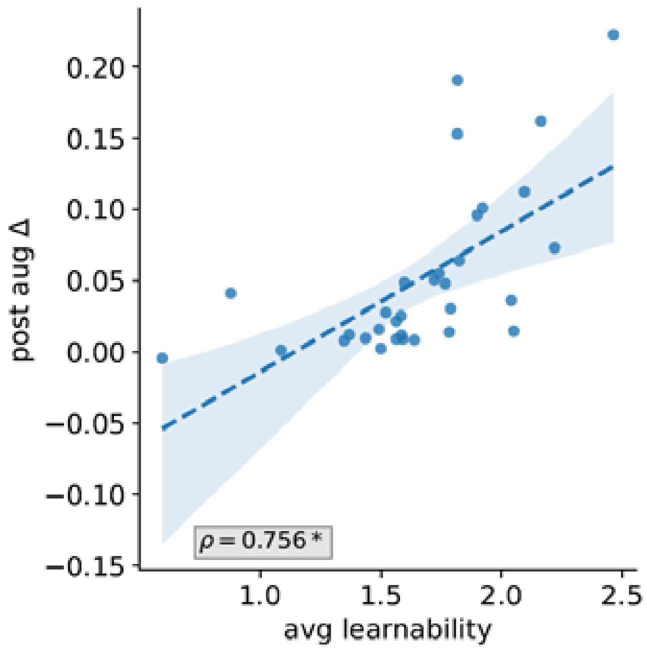
- We observe a negative correlation between learnability and robustness across all three datasets, validating **Hypothesis 1**.

> If a certain perturbation is more learnable for a model, the model will be less robust to this perturbation during test time.

Results

- Data augmentation with a perturbation the model is less robust to has more improvement on robustness (**Hypothesis 2**).

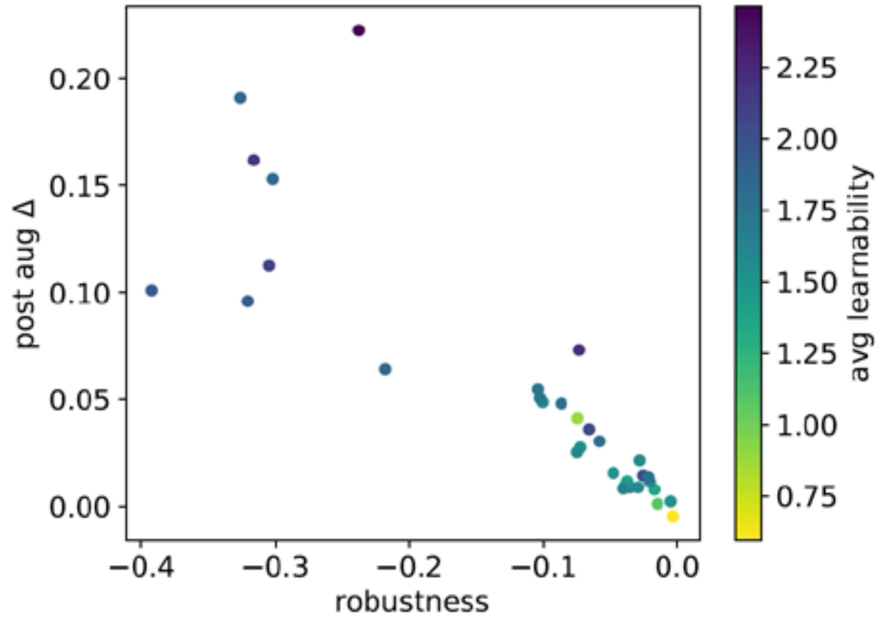
ρ	IMDB	YELP	QQP
Avg. learnability vs. robustness	-0.643*	-0.821*	-0.695*
Avg. learnability vs. post aug Δ	$\boxed{0.756*}$	0.846*	0.750*



(b) Learnability vs. Post Aug Δ

Results

- Data augmentation is *only* more effective at improving robustness against perturbations that a model is more sensitive to!



(c) Learn. vs. Robu. vs. Post Aug Δ

Conclusion

- We quantify how well the NLP model learns a perturbation with the learnability, which is grounded in the causality framework.
- We show a statistically significant inverse correlation between learnability and robustness.
- We provide an empirical explanation for why NLP models are less robust to some perturbations than others.