Competency Problems: On Finding and Removing Artifacts in Language Data

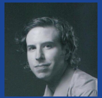

Matt Gardner*

William Merrill*

Jesse Dodge


Matthew E. Peters

Alexis Ross

Sameer Singh

Noah A. Smith

*Equal contribution

Motivation

Much work has shown many NLP datasets suffer from artifacts

- How to find artifacts?
- How to remove them?

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Table 4: Top 5 words by PMI(*word*, *class*), along with the proportion of *class* training samples containing *word*. MultiNLI is abbreviated to MNLI.

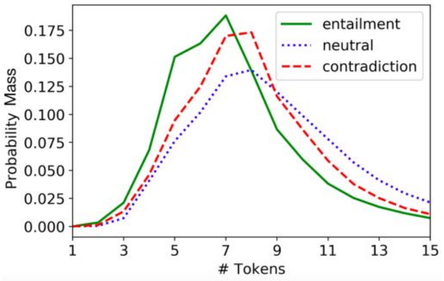


Figure 1: The probability mass function of the hypothesis length in SNLI, by class.

Gururangan et al., 2018

Heuristic	Supporting Cases	Contradicting Cases
Lexical overlap	2,158	261
Subsequence	1,274	72
Constituent	1,004	58

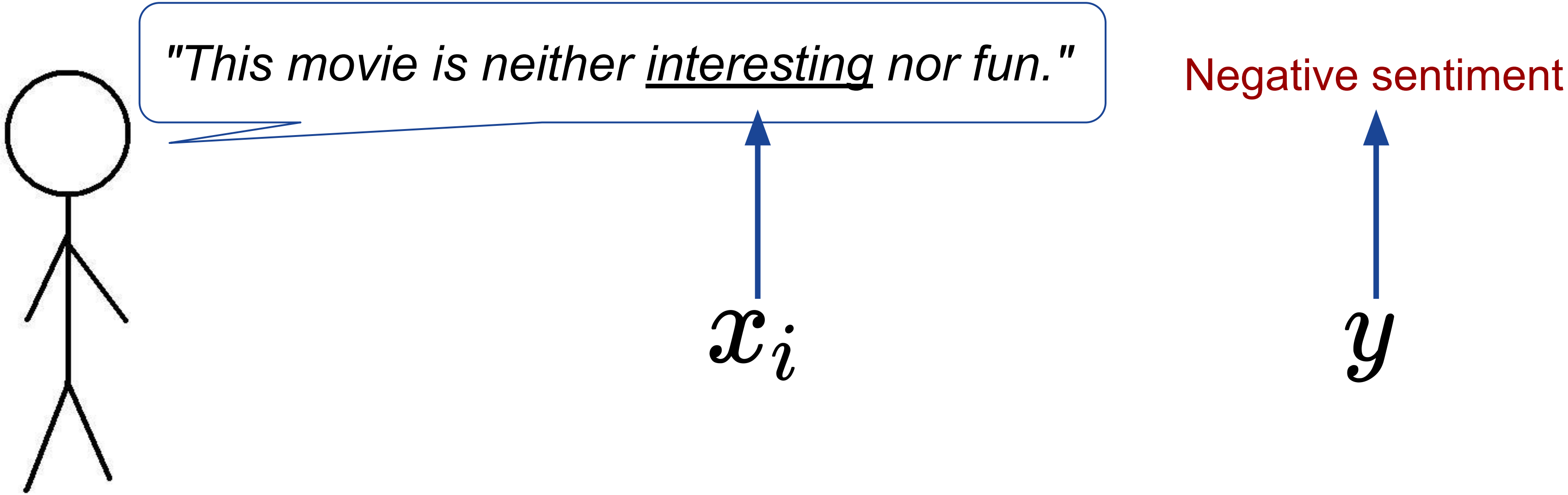
McCoy et al., 2019

Competency problems

Competency problem: all simple correlations between input features and output labels are spurious

$$p_u(y|x_i) = \frac{1}{|y|}$$

Example: sentiment analysis

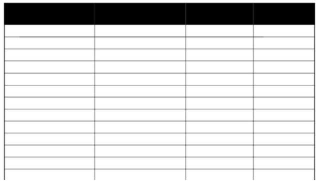


Even though $p(\text{negative} \mid \textit{interesting}) < .5$

Important clarification

We are not saying: natural data satisfies competency assumption (it doesn't)

!=



We are saying: generalizing to competency setting is **necessary** for NLU

∴ Competency is a good target for evaluating NLU systems

Finding dataset artifacts

Competency problems: measuring bias

Null hypothesis $p_u(y|x_i) = \frac{1}{|y|}$

Measured probabilities $\hat{p}_b(y|x_i)$

Concrete example: SNLI

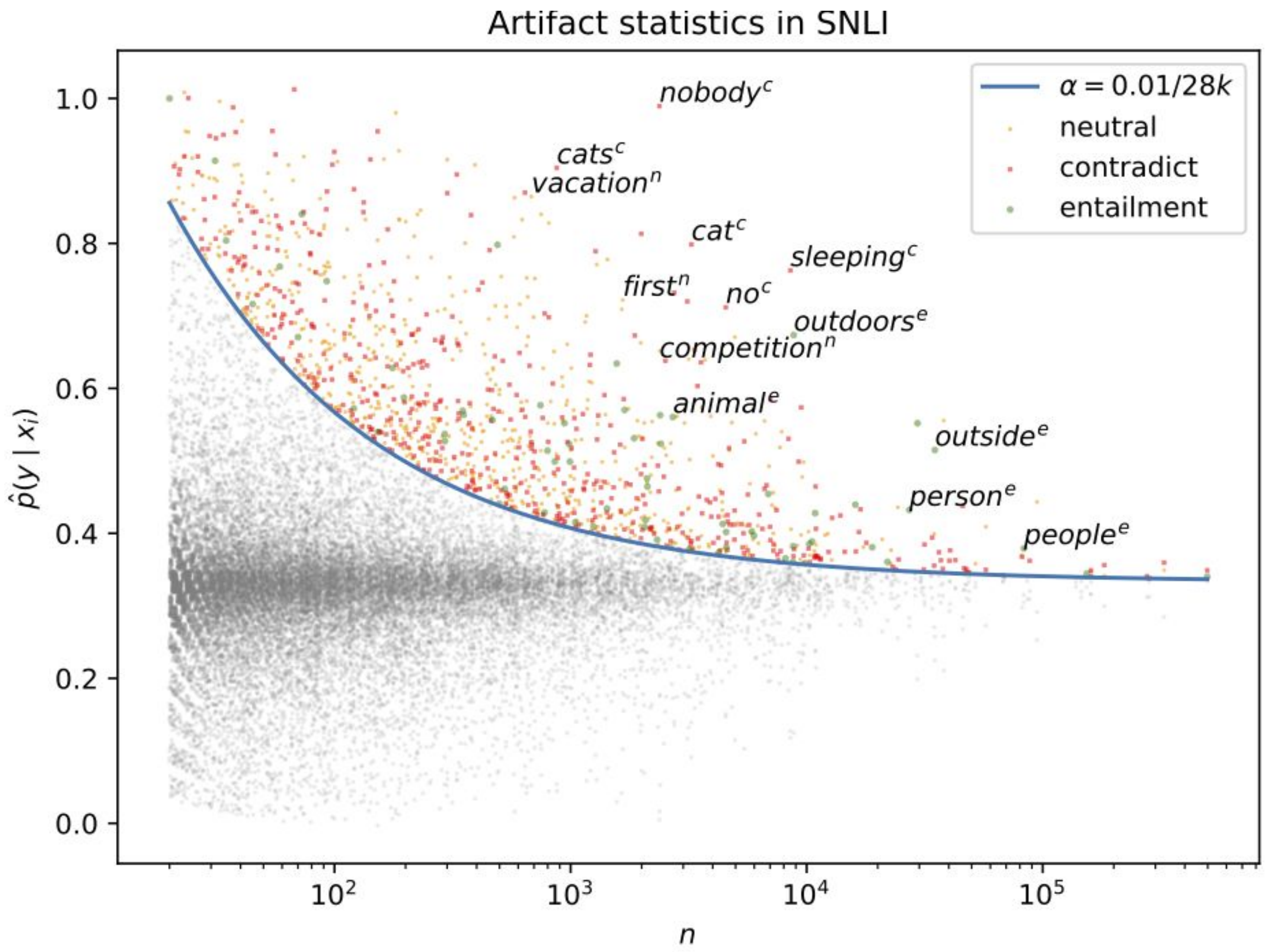
"Children smiling and waving at camera"
"There are children present"

entailment

x_i is indicator for the presence of a token, e.g., "children"

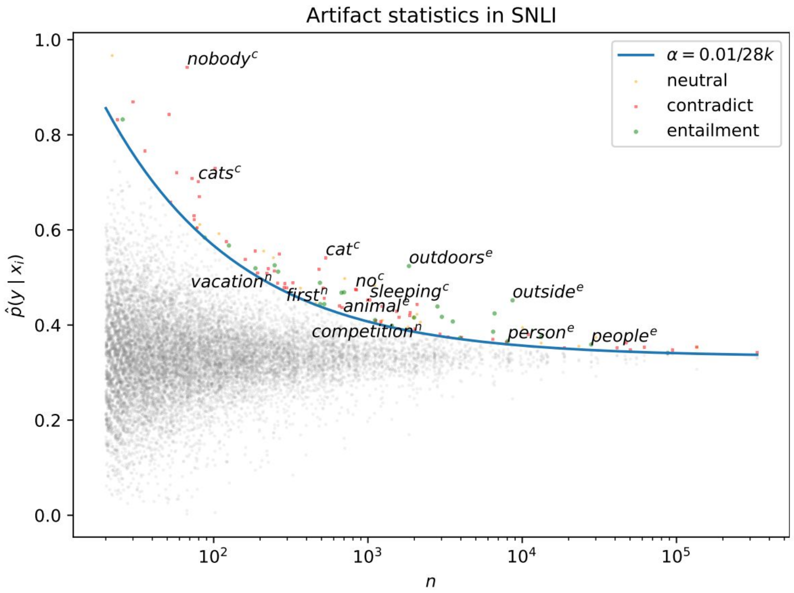
$y \in \{$ entailment, neutral, contradiction $\}$

Hypothesis test: $p_b(y|x_i) > \frac{1}{3}$

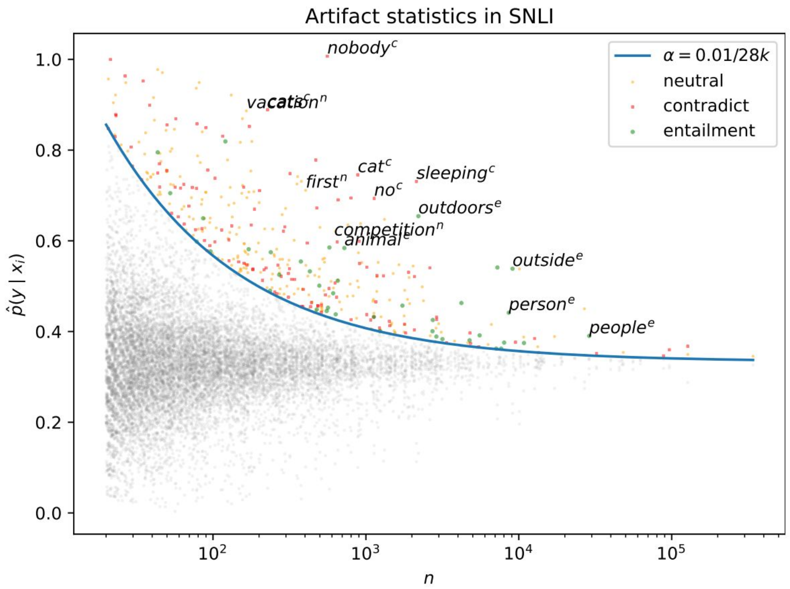


artifacts from Gururangan et al., 2018

Data harder for models is closer to a competency problem



Ambiguous instances
using dataset cartography (Swayamdipta et al., 2020)



Normal instances
with dataset size controlled to match ambiguous instances

Removing dataset artifacts

What can we do about bias in our data?

Local Edits

1. Randomly sample an instance $\mathbf{x}$ from a dataset $\mathcal{D}_b$ of n instances created under the biased distribution p_b.

2. Make some changes to $\mathbf{x}$ to arrive at $\mathbf{x}'$.

3. Manually label y' and add $\langle \mathbf{x}', y' \rangle$ to $\mathcal{D}_e$.

Edit sensitivity

With what probability does the edited label y' flip from y?

Local editing removes artifacts under the right conditions

Theorem.

$$\text{Edit sensitivity} = ½ \Rightarrow \left\{ p_e(y' \mid x_i') = ½ \right\}$$

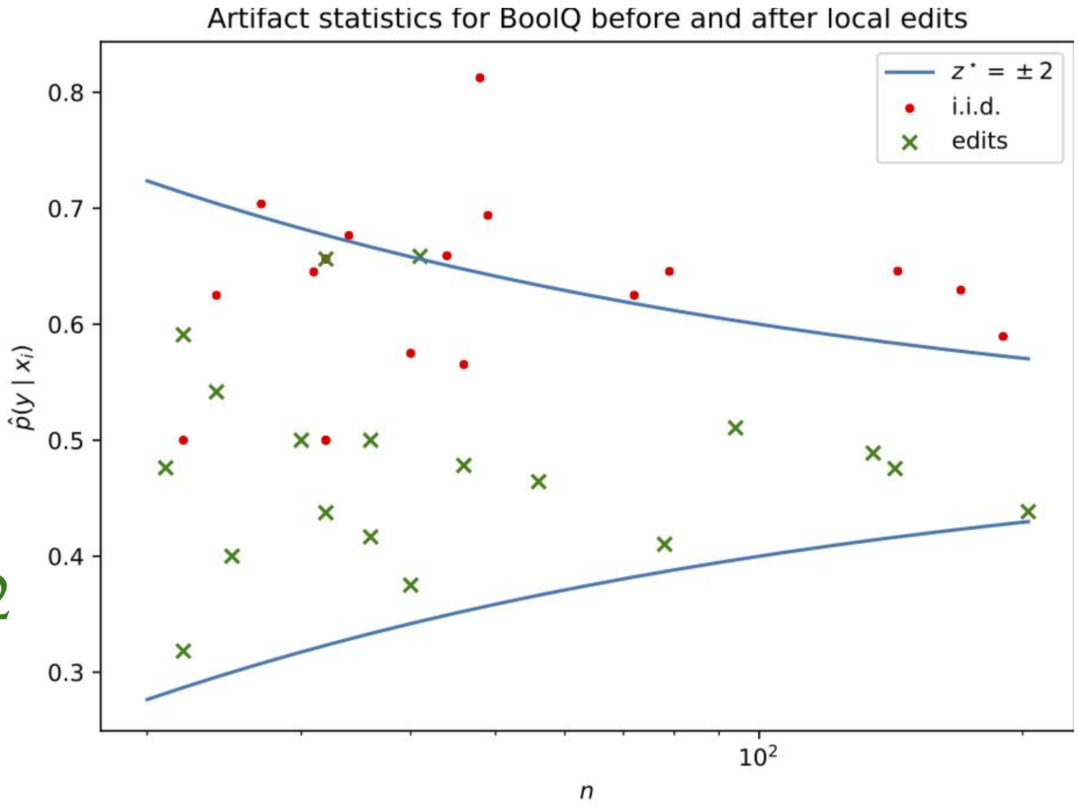
Edited data reflects competency

Local edits remove BoolQ artifacts

BoolQ: Boolean QA dataset
Clark et al., 2019

Gardner et al., 2020 created local edit version

✔Edit sensitivity = 0.52

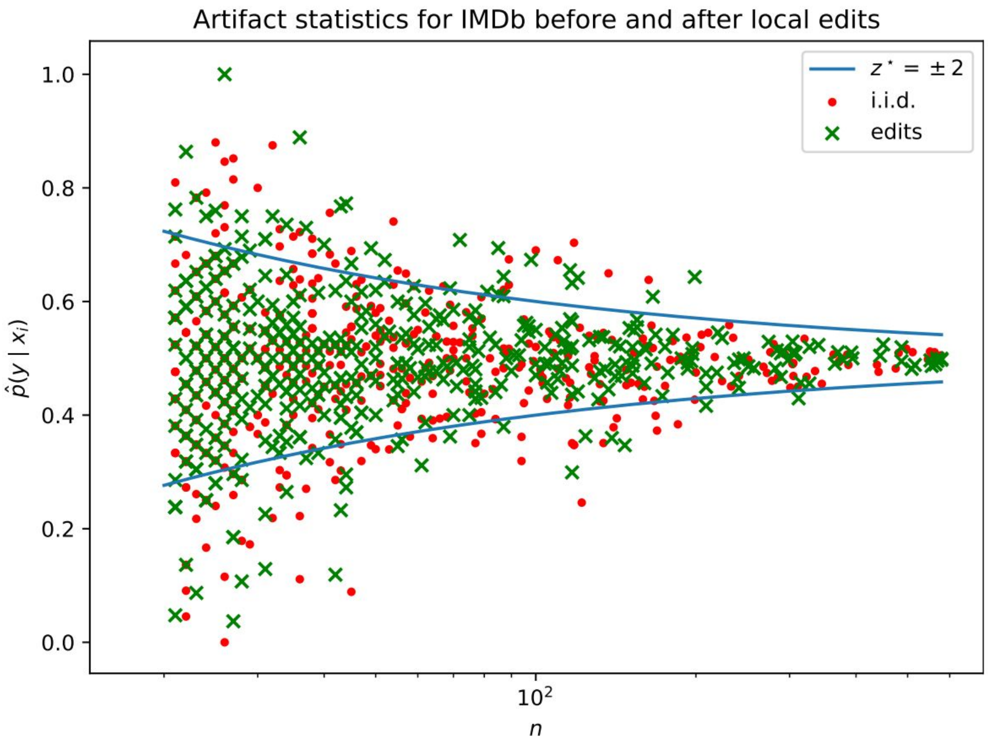


Local edits don’t remove IMDb artifacts

IMDb sentiment classification

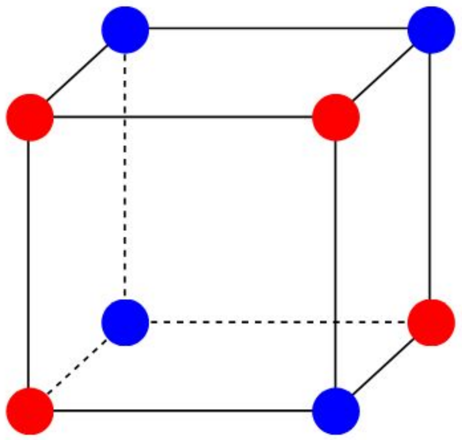
Gardner et al., 2020 created local edit version

❌Edit sensitivity = 1.00



Easter egg: boolean sensitivity

- No time here, but we also discuss connections between local editing and sensitivity in the theory of boolean functions
- Check Section 5



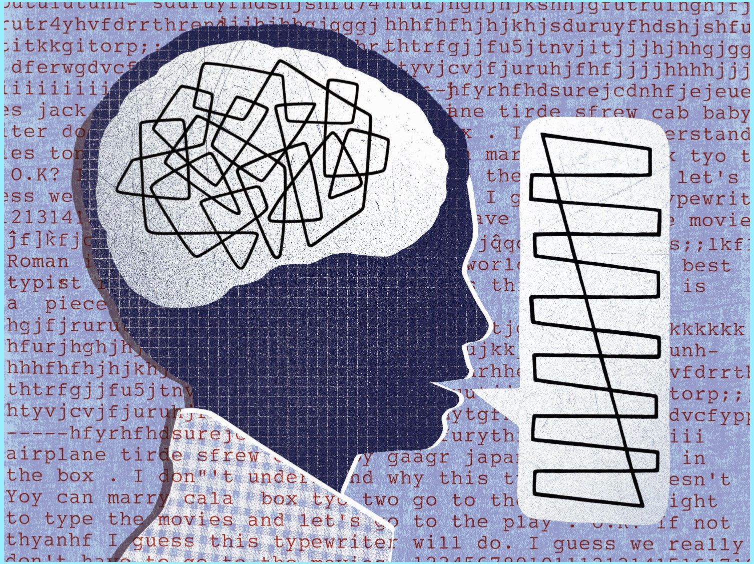
Value of competency problems framework

1. Statistical **test for artifacts** in a dataset
    a. Models are impacted negatively by these artifacts

2. Local edits algorithm to **remove artifacts** from datasets (with theoretical guarantees)
    a. Insight into how to design local editing procedure

Thank you!

Thank you!



AllenNLP

THE UNIVERSITY OF CALIFORNIA · IRVINE · LET THERE BE LIGHT

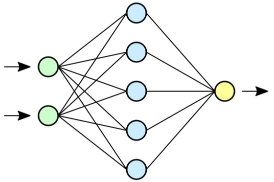
Ai2

Models learn these biases

Synthetic experiment: input single token to trained model

Occurs **more**
w/ contradict

"nobody"

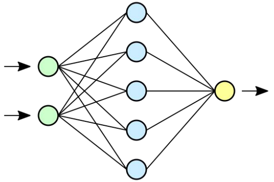


$\hat{p}_y$

vs

Occurs **less**
w/ contradict

"outside"



$\hat{p}'_y$

$$\hat{p}_y - \hat{p}'_y = \Delta\hat{p}_y$$

Models learn biases

Class	$\Delta\hat{p}_y$
entailment	+14.7 %
neutral	+7.9 %
contradiction	+12.5 %

Example: sentiment analysis

- For improving practical performance on narrow-domain sentiment analysis system, relying on correlation between “interesting” and negative sentiment is okay

- For understanding language like a human, single features should not be informative about the label!