On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations

Roy Schwartz
The Hebrew University of Jerusalem
TAU-NLP Seminar, 04/2022

Benchmarks in NLP

Natural Questions

A Benchmark for Question Answering Research.

SWAG

DROP

QuAC

Question Answering in Context

MultiNLI

SuperGLUE

SQuAD2.0

The Stanford Question Answering Dataset

Benchmarks in NLP

The Premise

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers♠ **Yonatan Bisk**♠ **Roy Schwartz**♠♡ **Yejin Choi**♠♡
♠Paul G. Allen School of Computer Science & Engineering, University of Washington
♡Allen Institute for Artificial Intelligence
{rowanz,ybisk,roysch,yejin}@cs.washington.edu
https://rowanzellers.com/swag

2017). **First**, our dataset poses a new challenge of grounded commonsense inference that is easy for humans (88%) while hard for current state-of-the-art NLI models (<60%). **Second**, our pro-

Benchmarks in NLP
Reality

Benchmark	Baseline	Shortly after
SWAG (Zellers et al., 2018)	52%	86% (Devlin et al., 2018)
DROP (Dua et al., 2019)	47 F1	90 F1 (Chen et al., 2020)
HellaSWAG (Zellers et al., 2019)	47%	93% (He et al., 2020)
WinoGrande (Sakaguchi et al., 2020)	53% AUC	88% AUC (Raffel et al., 2020)

A (Naive?) Conclusion

Microsoft creates AI that can read a document and answer questions about it as well as a person

January 15, 2018 | Allison Linn

More Like this

More Like this

Spurious Correlations

In statistics, a spurious relationship or spurious correlation is a mathematical relationship in which ***two or more events or variables*** *are* ***associated*** *but not* ***causally related****, due to either* ***coincidence*** *or the presence of a* ***certain third, unseen factor****.* [*Wikipedia*]

Spurious Correlations and NLP Benchmarks

- Instead of **understanding** the text, machines pick up on these **correlations** from the training data
  - They use the learned correlations to excel on the test sets

Spurious Correlations and NLP Benchmarks

- Instead of **understanding** the text, machines pick up on these **correlations** from the training data
  - They use the learned correlations to excel on the test sets
- This artificially **inflate** the **state of the art**

Spurious Correlations and NLP Benchmarks

- Instead of **understanding** the text, machines pick up on these **correlations** from the training data
  - They use the learned correlations to excel on the test sets
- This artificially **inflate** the **state of the art**
- As a result, many efforts exist to **mitigate these correlations**

Outline

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

Outline

- Background
    - Spurious correlations in NLP datasets
    - What makes a correlation spurious?
    - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
    - Practical and conceptual limitations
- Alternatives to dataset balancing
    - Richer context
    - Interactivity and abstention
    - Large-scale finetuning -> zero-/few-shot learning

Spurious Correlations in Vision and Language

- VQA dataset
  - Antol et al. (2015)
- Input: an image and a question
  - What sport is this man playing?
  - Do you see a shadow?
- Output: answer
  - Tennis, yes

Spurious Correlations in VQA

- 40% of the questions in VQA starting with "***What sport is this***" are answered with "***tennis***"

- "***yes***" is the answer to 87% of the questions in the VQA dataset starting with "***Do you see a***"
    - Zhang et al. (2016); Goyal et al. (2017)

ROC Story Cloze Task

Mostafazadeh et al. (2016)

Context	Right Ending	Wrong Ending
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.

- A story comprehension task

ROC Story Cloze Task

Mostafazadeh et al. (2016)

Context	Right Ending	Wrong Ending
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.

- A story comprehension task
- The task: given a story prefix, distinguish between the coherent and the incoherent endings

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Model	**Acc.**
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Model	Acc.
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Model	Acc.
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Right	Weight	Freq.	***Wrong***	Weight	Freq.
'ed .'	0.17	6.5%	START NNP	0.21	54.8%
'and '	0.15	13.6%	NN .	0.17	47.5%
JJ	0.14	45.8%	NN NN .	0.15	5.1%
to VB	0.13	20.1%	VBG	0.11	10.1%
'd th'	0.12	10.9%	START NNP VBD	0.11	41.9%

Spurious Correlations in ROC

S. et al. (2017); Cai et al. (2017)

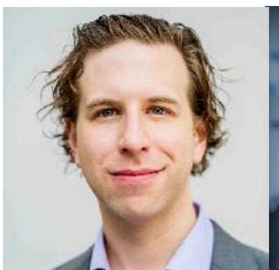
- Train a binary classifier on **the endings only**
  - Ignoring the story prefix

Right Ending	Wrong Ending
Tom asked Sheryl to marry him.	He wiped mud off of his boot.

Model	Acc.
DSSM (Mostafazadeh et al., 2016a)	0.585
ukp (Bugert et al., 2017)	0.717
tbmihaylov (Mihaylov and Frank, 2017)	0.724
†EndingsOnly (Cai et al., 2017)	0.725
cogcomp	0.744
HIER,ENCPLOTEND,ATT (Cai et al., 2017)	0.747
RNN	0.677
†Ours	0.724
Combined (ours + RNN)	**0.752**
Human judgment	1.000

Right	Weight	Freq.	***Wrong***	Weight	Freq.
'ed .'	0.17	6.5%	START NNP	0.21	54.8%
'and '	0.15	13.6%	NN .	0.17	47.5%
JJ	0.14	45.8%	NN NN .	0.15	5.1%
to VB	0.13	20.1%	VBG	0.11	10.1%
'd th'	0.12	10.9%	START NNP VBD	0.11	41.9%






SNLI and MNLI

SNLI and MNLI

What about NLI datasets?

SNLI and MNLI

Great question!

What about NLI datasets?

SNLI and MNLI

Great question!

What about NLI datasets?

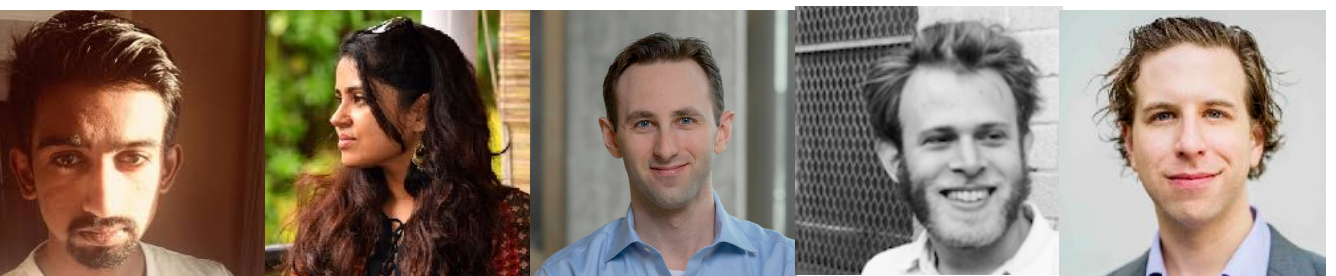
Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

SNLI (Bowman et al., 2015); MNLI (Williams et al., 2018)

Spurious Correlations in NLI Datasets

Gururangan, Swaymdipta, Levy, S., Bowman, Smith (2018); Poliak et al. (2018); Tsuchiya (2018)

- Train a hypothesis-only classifier
  - No premise

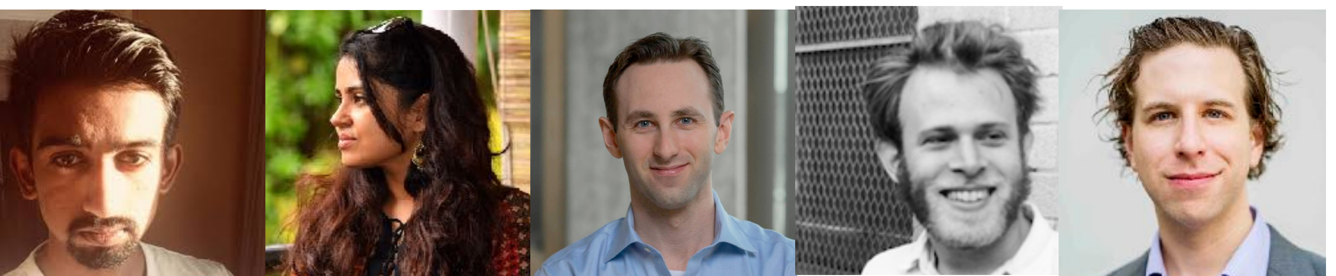


Spurious Correlations in NLI Datasets

Gururangan, Swaymdipta, Levy, S., Bowman, Smith (2018); Poliak et al. (2018); Tsuchiya (2018)

- Train a hypothesis-only classifier
  - No premise

Model	SNLI	MultiNLI	
		Matched	Mismatched
majority class	34.3	35.4	35.2
fastText	**67.0**	**53.9**	**52.3**



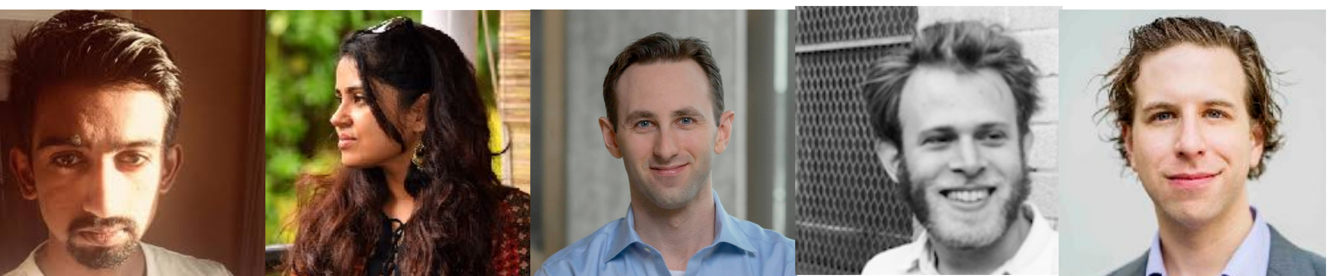
Spurious Correlations in NLI Datasets

Gururangan, Swaymdipta, Levy, S., Bowman, Smith (2018); Poliak et al. (2018); Tsuchiya (2018)

- Train a hypothesis-only classifier
  - No premise

Model	SNLI	MultiNLI Matched	MultiNLI Mismatched
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

	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%



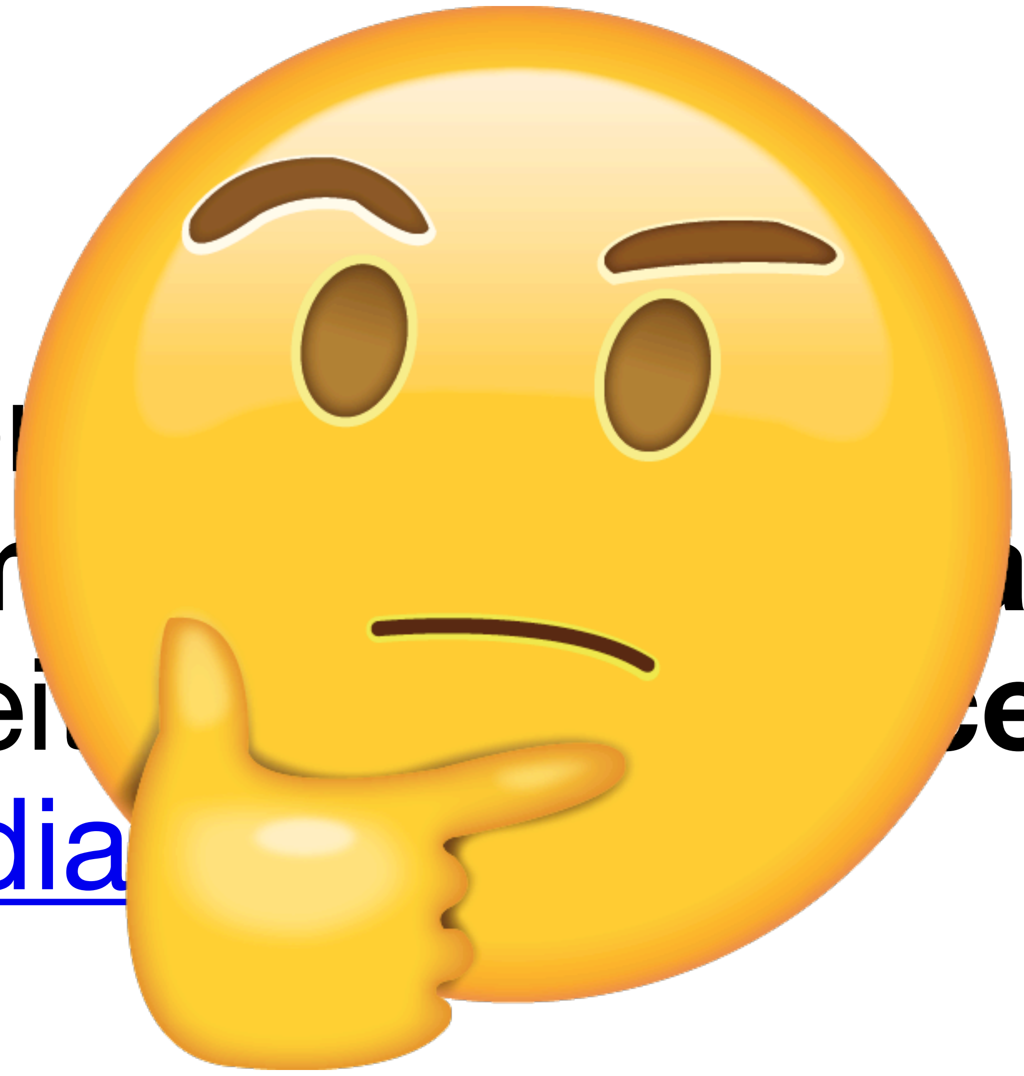
Other Spurious Correlations

- Other tasks
  - Question answering (Kaushik & Lipton, 2018)
  - Winograd Schema (Elazar et al., 2021)
- Are We Modeling the Task or the Annotator?
  - Geva et al. (2019)

What are Spurious Correlations?

- In statistics, a spurious relationship or spurious correlation is a mathematical relationship in which **two or more events or variables** are **associated** but not **causally related**, due to either **coincidence** or the presence of a **certain third, unseen factor**. Wikipedia

What are Spurious Correlations?



- In statistics, a spurious relatio correlation is a mathematical relationship in which **two or m** **ariables** are **associated** but not **causally related**, due to ei **e** or the presence of a **certain third, unseen factor**. [Wikipedia]

What are Spurious Correlations?

Ingenuine correlations

- A feature correlated with some output label for no apparent reason
  - E.g., “cat” and “sleeping” are correlated with contradictions in SNLI (Gururangan et al., 2018)
  - Wang and Culotta, 2020; Rogers, 2021

What are Spurious Correlations?

Ingenuine **correlations**

- A feature correlated with some output label for no apparent reason
  - E.g., “cat” and “sleeping” are correlated with contradictions in SNLI (Gururangan et al., 2018)
  - Wang and Culotta, 2020; Rogers, 2021
- An appealing definition

What are Spurious Correlations?

Ingenuine correlations

- A feature correlated with some output label for no apparent reason
  - E.g., "cat" and "sleeping" are correlated with contradictions in SNLI (Gururangan et al., 2018)
  - Wang and Culotta, 2020; Rogers, 2021
- An appealing definition
- But somewhat subjective
  - E.g., the word "not" indicating NLI contradictions; "amazing" as a feature for positive sentiment

What are Spurious Correlations?

Ungeneralizable correlations

- A feature that works well for specific examples but **does not hold in general**
    - Chang et al., 2021; Yaghoobzadeh et al., 2021

What are Spurious Correlations?

Ungeneralizable correlations

- A feature that works well for specific examples but **does not hold in general**
  - Chang et al., 2021; Yaghoobzadeh et al., 2021
- Does not address the nature of the feature
  - Whether genuine or not

What are Spurious Correlations?

Ungeneralizable correlations

- A feature that works well for specific examples but **does not hold in general**
    - Chang et al., 2021; Yaghoobzadeh et al., 2021
- Does not address the nature of the feature
    - Whether genuine or not
- But does assume the feature is *important*
    - And thus somewhat subjective

What are Spurious Correlations?

every-word

- *Every* simple correlation between single word features and output labels is spurious
  - Gardner et al., 2021

What are Spurious Correlations?

every-word

- *Every* simple correlation between single word features and output labels is spurious
    - Gardner et al., 2021
- *Competent* datasets: the marginal probability for every feature is uniform over the class label
    - $\forall x_i, y \in Y, p(y|x_i) = \frac{1}{|Y|}$

Mitigating Spurious Correlations

- Change the model
  - Adversarial networks (Belinkov et al., 2019; Grand and Belinkov, 2019; Wang et al., 2019; Cadene et al., 2019)
  - Model ensembles (Clark et al., 2019,2020; He et al., 2019; Bahng et al., 2020)

Mitigating Spurious Correlations

- Change the model
  - Adversarial networks (Belinkov et al., 2019; Grand and Belinkov, 2019; Wang et al., 2019; Cadene et al., 2019)
  - Model ensembles (Clark et al., 2019,2020; He et al., 2019; Bahng et al., 2020)
- Change the data
  - **Data balancing**

Mitigating Spurious Correlations via Dataset Balancing

Augmentation

- The key idea: balance-out spurious correlations
- Vision and Language datasets
  - VQA 2.0 (Goyal et al. ,2017)
  - GQA (Hudson and Manning, 2019)
- Language only
  - ROC stories cloze task 1.5 (Sharma et al., 2018)

Mitigating Spurious Correlations via Dataset Balancing

Augmentation

- The key idea: balance-out spurious correlations
- Vision and Language datasets
  - VQA 2.0 (Goyal et al. ,2017)
  - GQA (Hudson and Manning, 2019)
- Language only
  - ROC stories cloze task 1.5 (Sharma et al., 2018)

Mitigating Spurious Correlations via Dataset Balancing

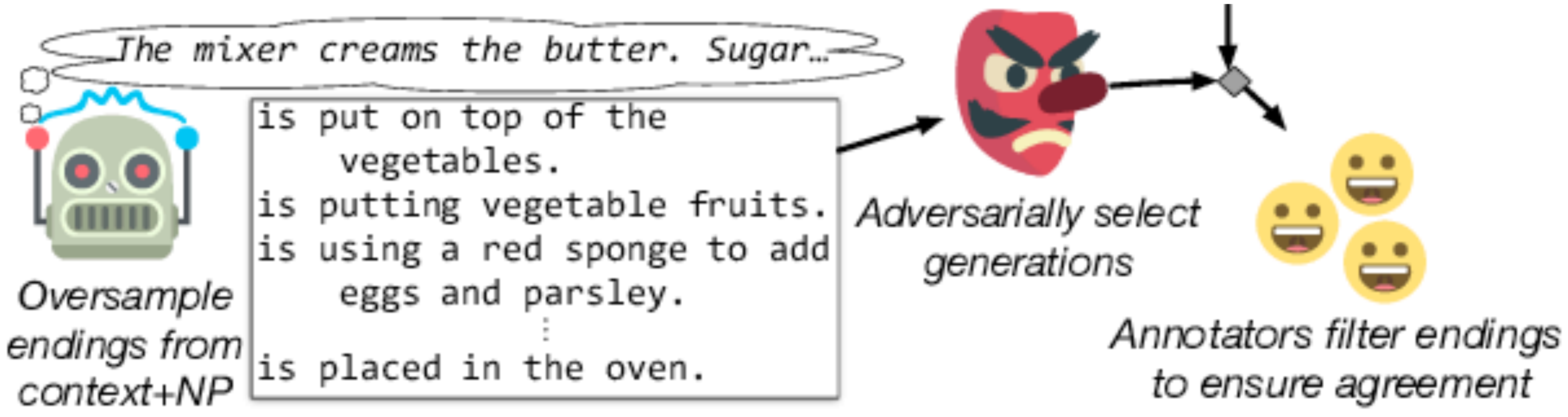
Filtering

- Adversarial filtering
  - Zellers, Bisk, S., Choi (2018)

Mitigating Spurious Correlations via Dataset Balancing

Filtering

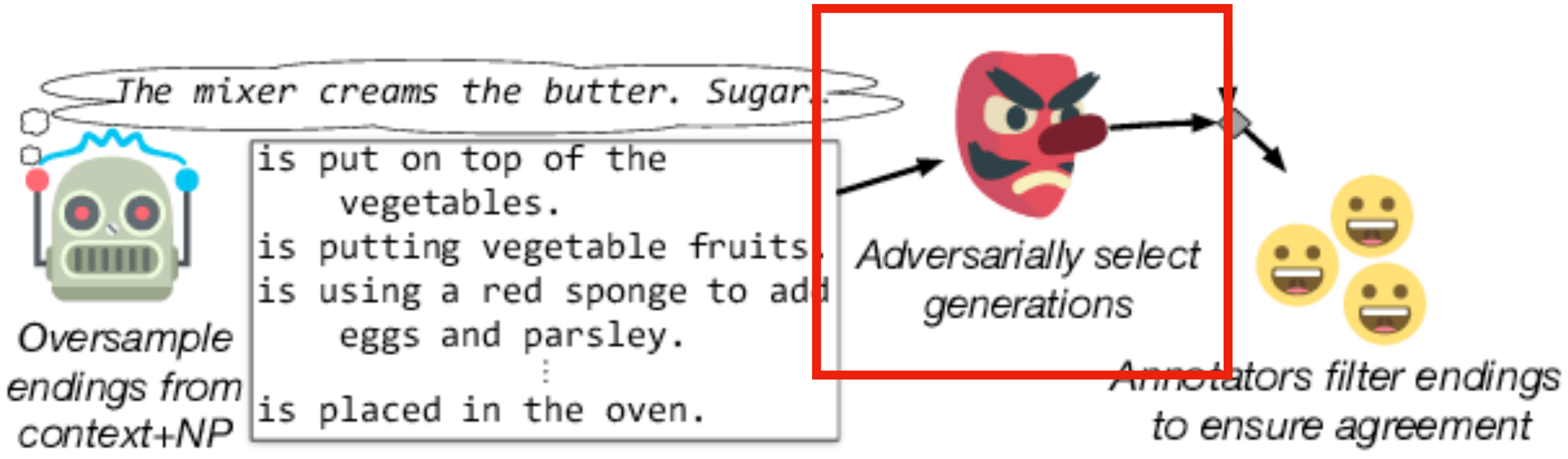
- Adversarial filtering
    - Zellers, Bisk, S., Choi (2018)



Mitigating Spurious Correlations via Dataset Balancing

Filtering

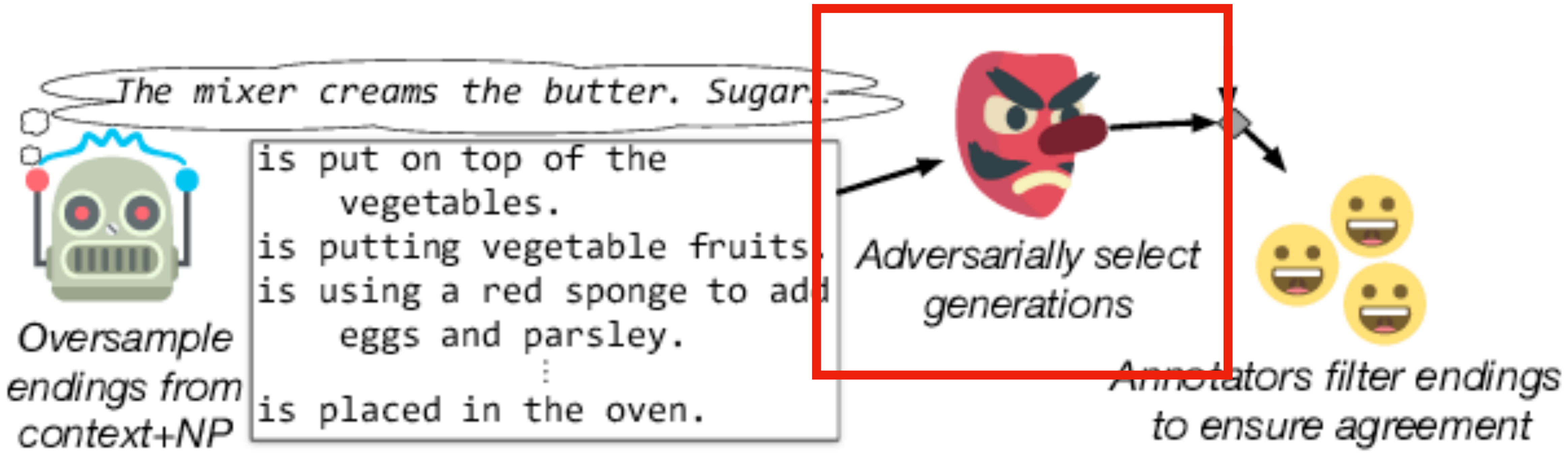
- Adversarial filtering
  - Zellers, Bisk, S., Choi (2018)



Mitigating Spurious Correlations via Dataset Balancing

Filtering

- Adversarial filtering
  - Zellers, Bisk, S., Choi (2018)
- Designed to "***systematically discover and filter any dataset artifact*** *in crowd- sourced commonsense problems*" (Le Bras et al., 2020)



Filtering as Balancing

- As the adversarial model grows, models will pick up subtler correlations
  - Resulting in a fully *balanced* dataset

Filtering as Balancing

- As the adversarial model grows, models will pick up subtler correlations
  - Resulting in a fully *balanced* dataset
- Widely adopted
  - Record (Zhang et al., 2018)
  - DROP (Dua et al., 2019)
  - HellaSWAG (Zellers et al., 2019)
  - αNLI (Bhagavatula et al., 2019)
  - WinoGrande (Sakaguchi et al., 2020)
  - …

Outline

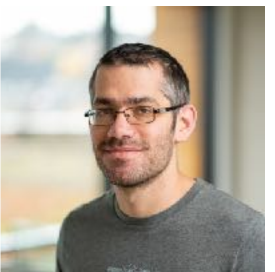
- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

Outline

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations

Roy Schwartz **Gabriel Stanovsky**
School of Computer Science, The Hebrew University of Jerusalem
{roy.schwartz1,gabriel.stanovsky}@mail.huji.ac.il



Balancing too Little is Insufficient

Toy Example

Split	Text	Label
Train	very good	+
	very bad	−
	not good	−
	not bad	+
Test	not very good	−
	good	+

Balancing too Little is Insufficient

Toy Example

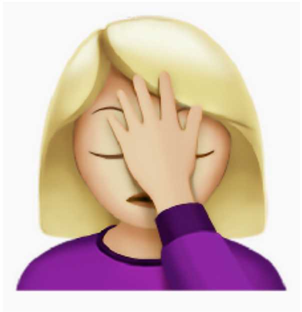

The dataset is balanced for unigrams

Split	Text	Label
Train	very good	+
	very bad	−
	not good	−
	not bad	+
Test	not very good	−
	good	+

Balancing too Little is Insufficient

Toy Example

The dataset is balanced for unigrams



But still contains spurious **bigrams** features

- E.g., "*very good*", as "***not** very good*" yields negative sentiment

Split	Text	Label
Train	very good	+
	very bad	−
	not good	−
	not bad	+
Test	not very good	−
	good	+

Balancing too Little is Insufficient

Natural Language

- The same example can apply with larger *n*’s

Balancing too Little is Insufficient

Natural Language

- The same example can apply with larger *n*’s
- More broadly, any phrase or feature combination can alter its meaning in some context
    - Negation, sarcasm, humor, …

Balancing too Little is Insufficient

Natural Language

- The same example can apply with larger *n*'s
- More broadly, any phrase or feature combination can alter its meaning in some context
    - Negation, sarcasm, humor, …
- As a result, balancing too little is **insufficient** for mitigating all spurious correlations

Too much Balancing Leaves Nothing

Toy Example

Original Train Set	
Input	**Label**
0 0	0
0 1	1
1 0	1
1 1	0

Too much Balancing Leaves Nothing

Toy Example

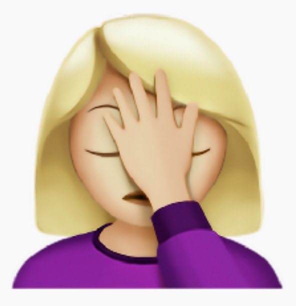
The dataset is also balanced for unigrams

Original Train Set	
Input	**Label**
0 0	0
0 1	1
1 0	1
1 1	0

Too much Balancing Leaves Nothing

Toy Example

The dataset is also balanced for unigrams



But if we balance it for bigrams, we are left with **no learnable signal**

Original Train Set		Augmented Samples	
Input	**Label**	**Input**	**Label**
0 0	0	*0 0	1
0 1	1	*0 1	0
1 0	1	*1 0	0
1 1	0	*1 1	1

Too much Balancing Leaves Nothing

More Broadly

- Consider an NLP dataset *D* with maximal length *n*

Too much Balancing Leaves Nothing

More Broadly

- Consider an NLP dataset D with maximal length n
- By definition, balancing any combination of up to n features (including) leaves no learnable signal in D

Too much Balancing Leaves Nothing

More Broadly

- Consider an NLP dataset *D* with maximal length *n*
- By definition, balancing any combination of up to *n* features (including) leaves no learnable signal in *D*
- Conclusion: *balancing too much* is not helpful either

Does a sweet-spot exist between balancing too little and too much?

Is Balancing even Desired?

- Dataset balancing prevents models from having a fallback option in cases of uncertainty
  - As these would evidently cause it to make mistakes on some inputs

Is Balancing even Desired?

- Dataset balancing prevents models from having a fallback option in cases of uncertainty
  - As these would evidently cause it to make mistakes on some inputs
- But fallback meanings are crucial for language understanding, as contexts are often underspecified
  - Graesser, 2013

Is Balancing even Desired?

- Especially relevant for world knowledge and common-sense knowledge
  - Joe Biden is the president of the US
  - A person is typically happy when they receive a present

Who is the president of the U.S.?

Context	**Answer**
∅	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah “Jed” Bartlet

Is Balancing even Desired?

- Especially relevant for world knowledge and common-sense knowledge
  - Joe Biden is the president of the US
  - A person is typically happy when they receive a present
- As a result, dataset balancing is **undesired**

Who is the president of the U.S.?

Context	**Answer**
$\emptyset$	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah "Jed" Bartlet

Is dataset balancing the right way forward?

Outline

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

Outline

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

Augmenting Datasets with Rich Contexts

Current practice: *Dataset Balancing*

- Instead of *unlearning* certain information, we should be focusing on learning and modeling **richer contexts**

Augmenting Datasets with Rich Contexts

Current practice: *Dataset Balancing*

- Instead of *unlearning* certain information, we should be focusing on learning and modeling **richer contexts**
- Example: negation
  - Instead of *unlearning* what "amazing" means, we should focus on learning what "**not** amazing" means
  - Negation still poses a challenge for modern NLP models (Hossain et al., 2020,2022)

Augmenting Datasets with Rich Contexts

More Details

- Other examples
  - Sarcasm (Davidov et al., 2010; Oprea and Magdy, 2020)
  - Humor (Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2020)
  - Metaphors (Tsvetkov et al., 2014; Mohammad et al., 2016)
  - More generally: broad coverage semantics (e.g., CCG, UCCA, AMR)

Augmenting Datasets with Rich Contexts

More Details

- Other examples
  - Sarcasm (Davidov et al., 2010; Oprea and Magdy, 2020)
  - Humor (Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2020)
  - Metaphors (Tsvetkov et al., 2014; Mohammad et al., 2016)
  - More generally: broad coverage semantics (e.g., CCG, UCCA, AMR)
- Concrete suggestions: adding documents with such contexts throughout the (pre)training corpus
  - Or alternatively, as a continued pretraining step to existing pretrained models

Abstention/Interaction

Motivation

To my great surprise, the movie turned out different than what I thought.

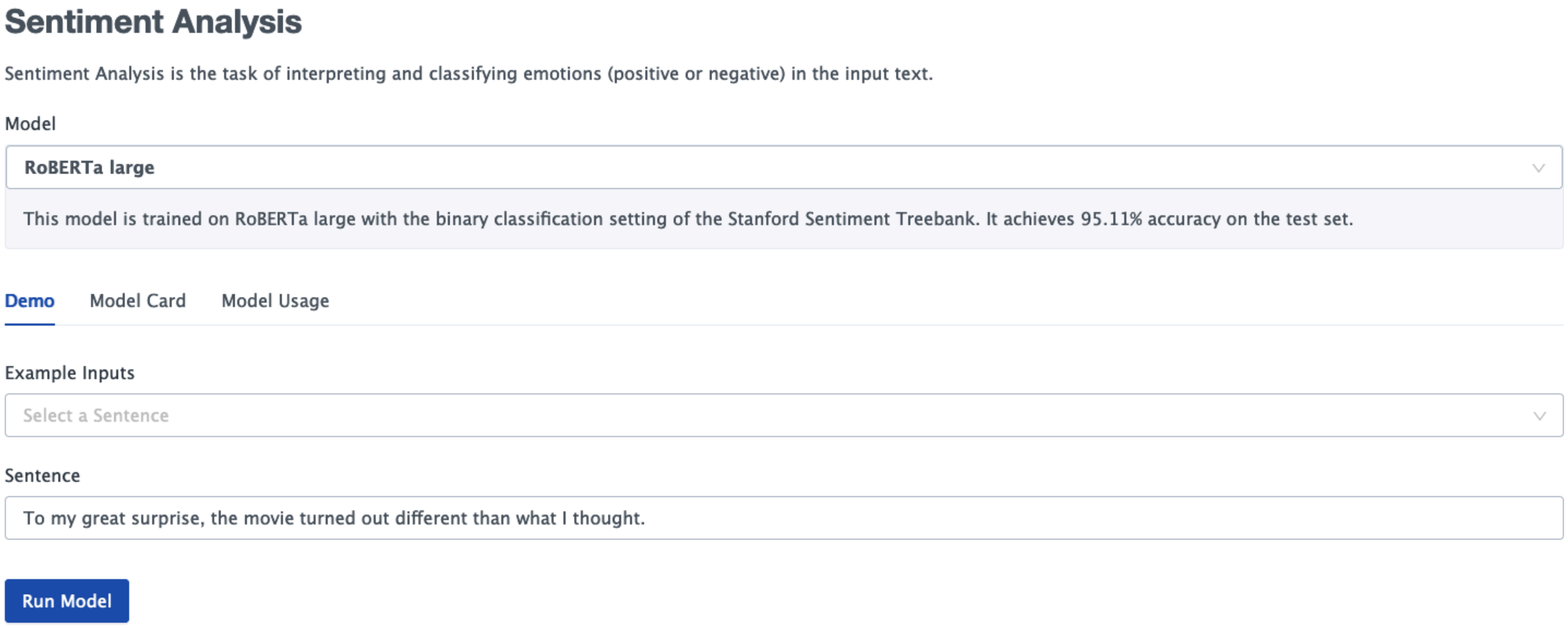
Abstention/Interaction

Motivation

To my $\boxed{\text{great surprise,}}$ the movie turned out different than what I thought.

Abstention/Interaction

Current practice: *a closed labeled set*



Abstention/Interaction

Current practice: *a closed labeled set*

Sentiment Analysis

Sentiment Analysis is the task of interpreting and classifying emotions (positive or negative) in the input text.

Model

RoBERTa large

This model is trained on RoBERTa large with the binary classification setting of the Stanford Sentiment Treebank. It achieves 95.11% accuracy on the test set.

Demo Model Card Model Usage

Example Inputs

Select a Sentence

Sentence

To my great surprise, the movie turned out different than what I thought.

Run Model

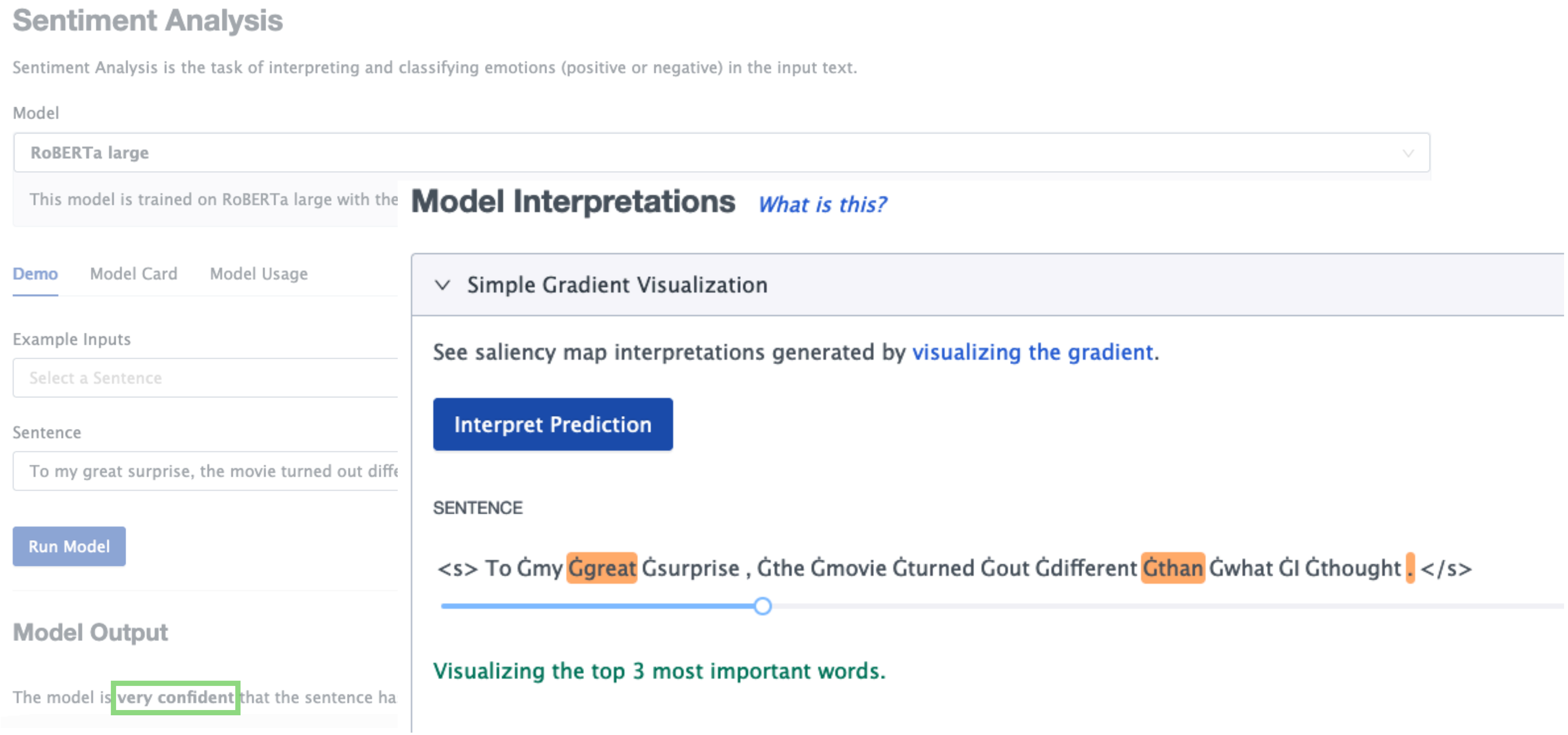
Model Output

Share

The model is **very confident** that the sentence has a **positive** sentiment.

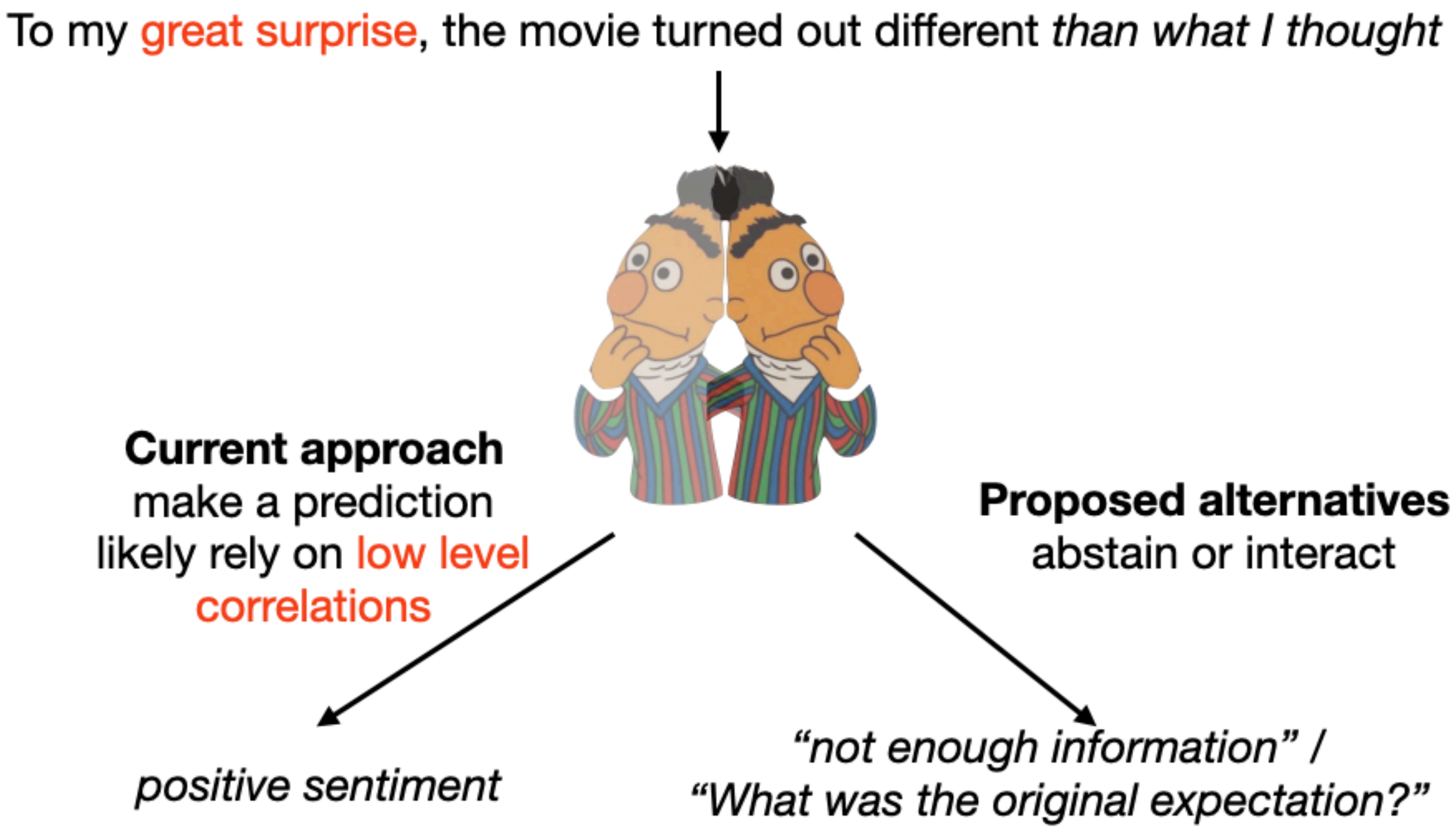
Abstention/Interaction

Current practice: *a closed labeled set*



Abstention/Interaction

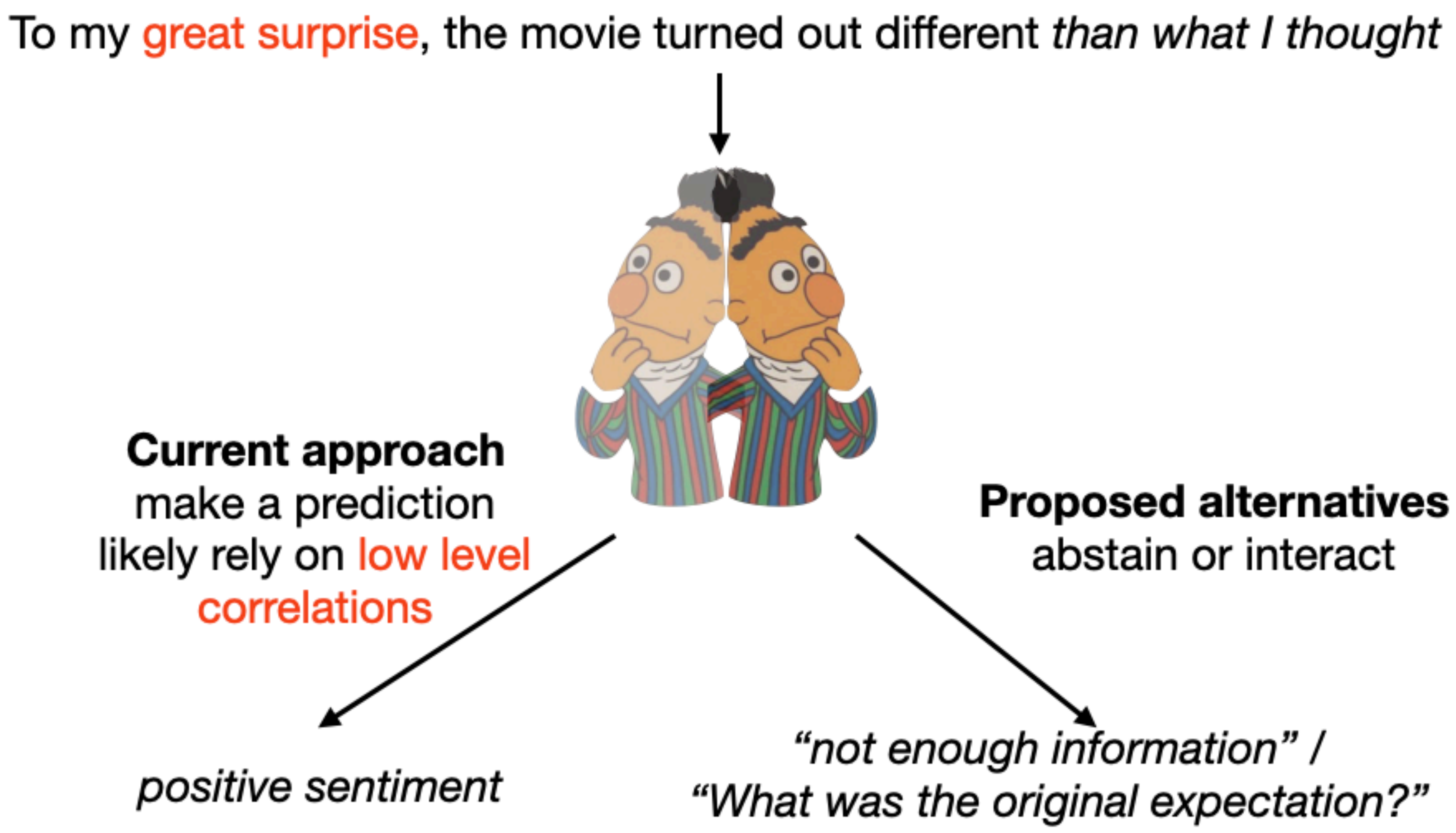
Proposal



Abstention/Interaction

Proposal

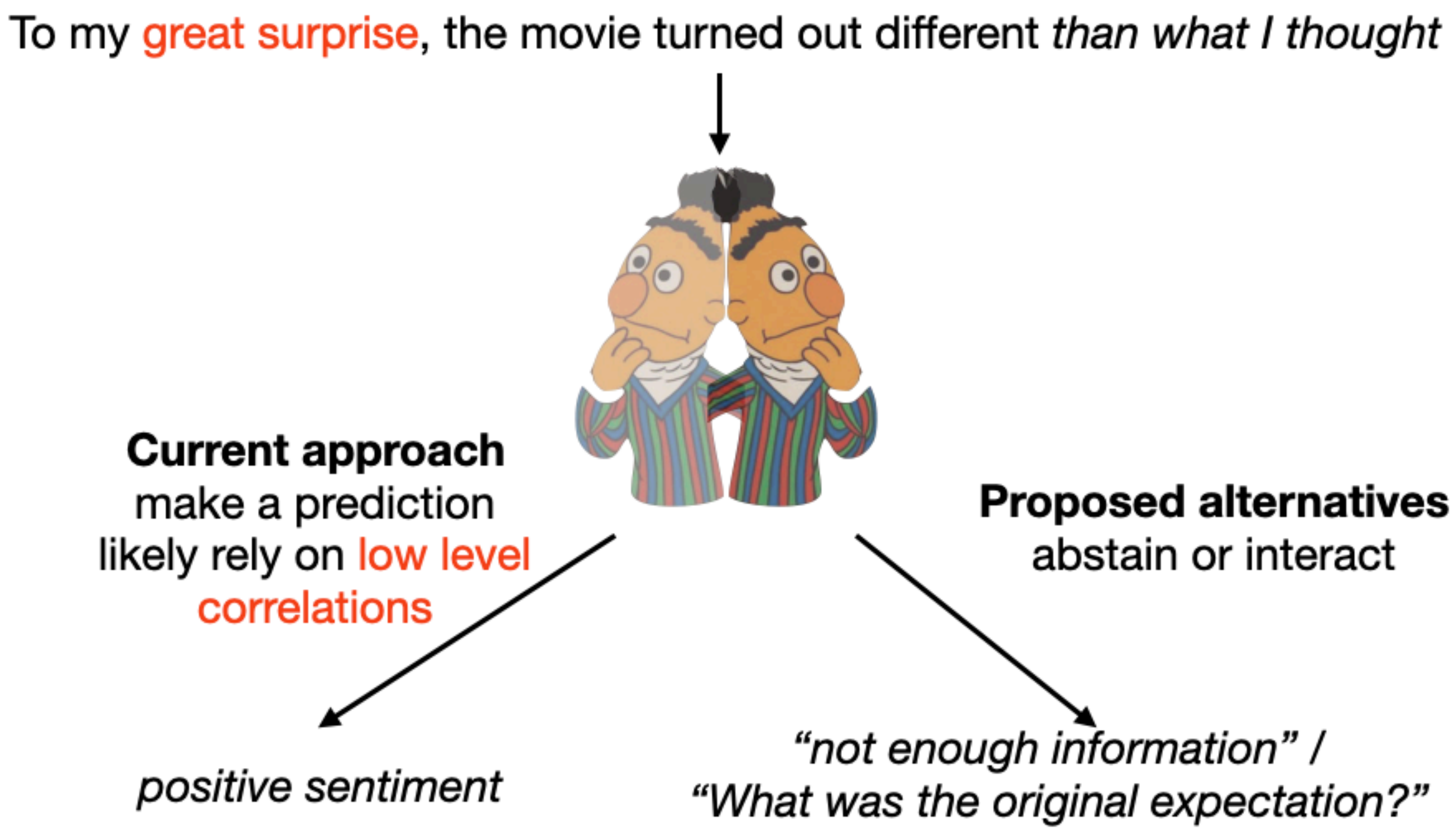
- Abstain / interact when models cannot make a confident decision
    - Chow, 1957; Hellman, 1970; Laidlaw and 525 Feizi, 2019; Balcan et al., 2020



Abstention/Interaction

Proposal

- Abstain / interact when models cannot make a confident decision
    - Chow, 1957; Hellman, 1970; Laidlaw and 525 Feizi, 2019; Balcan et al., 2020
- One example: datasets with unanswerable questions
    - Ray et al., 2016; Rajpurkar et al., 2018; Sulem et al., 2021



Few-shot Learning

Current Practice: Large-scale Fine-tuning

- Zero- and few-shot learning has improved dramatically
    - Sometimes reaching human-level performance (Schick and Schütze, 2021; Shin et al., 2020; Gu et al., 2021)

Few-shot Learning

Current Practice: Large-scale Fine-tuning

- Zero- and few-shot learning has improved dramatically
  - Sometimes reaching human-level performance (Schick and Schütze, 2021; Shin et al., 2020; Gu et al., 2021)
- One way to mitigate spurious correlations is to minimize manual annotation

Few-shot Learning

Current Practice: Large-scale Fine-tuning

- Zero- and few-shot learning has improved dramatically
    - Sometimes reaching human-level performance (Schick and Schütze, 2021; Shin et al., 2020; Gu et al., 2021)
- One way to mitigate spurious correlations is to minimize manual annotation
- **Do we still need large-scale fine-tuning?**

The End of Large-scale Fine-tuning?

- Limitations
  - Some spurious correlations may be picked up by the small number of examples
  - Or during pretraining (Gehman et al., 2020; Birhane et al., 2021; Dodge et al., 2021)
- Which tasks?
  - Large-scale supervision might still be necessary for some tasks (dialogue, summarization, …)
  - A rule of thumb: datasets or tasks for which the **state of the art** is close to or surpasses the **human baseline**

A Note on Social Biases

- Societal biases are often an undesired artifact of NLP models
  - E.g., gender, race
- In such cases, there might be a justification to *unlearn* them via dataset balancing
  - However, it is not clear that this is a practical goal

Summary

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning

Summary

- Background
  - Spurious correlations in NLP datasets
  - What makes a correlation spurious?
  - Mitigating spurious correlations via dataset balancing
- On the limitations of dataset balancing
  - Practical and conceptual limitations
- Alternatives to dataset balancing
  - Richer context
  - Interactivity and abstention
  - Large-scale finetuning -> zero-/few-shot learning