Foundation Models Can Robustify Themselves, For Free

Dyah Adila*, Changho Shin*, Linrong Cai, Frederic Sala

adila@wisc.edu, cshin23@wisc.edu

NeurIPS 2023 R0-FoMo workshop

Foundation Models Can **Robustify** Themselves, **For Free**

More **robust**

"**Great purchase**! If you want to ruin your weekend"

"Great purchase! **If** you want to ruin your weekend"

Free?

In The Ideal World..

1 Oracle

This is sarcasm, “great purchase” does not imply positive sentiment. Focus on the part after “if”

Image source: Amazon

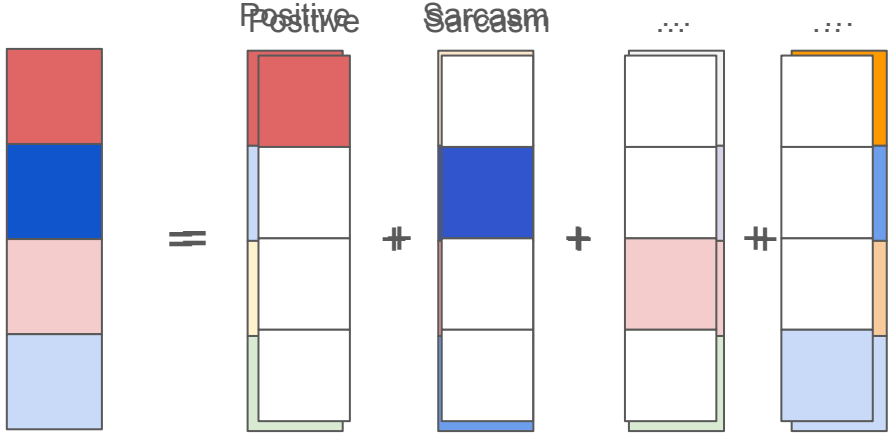
2 Disentangled representation

In Reality..

1 ~~Oracle~~

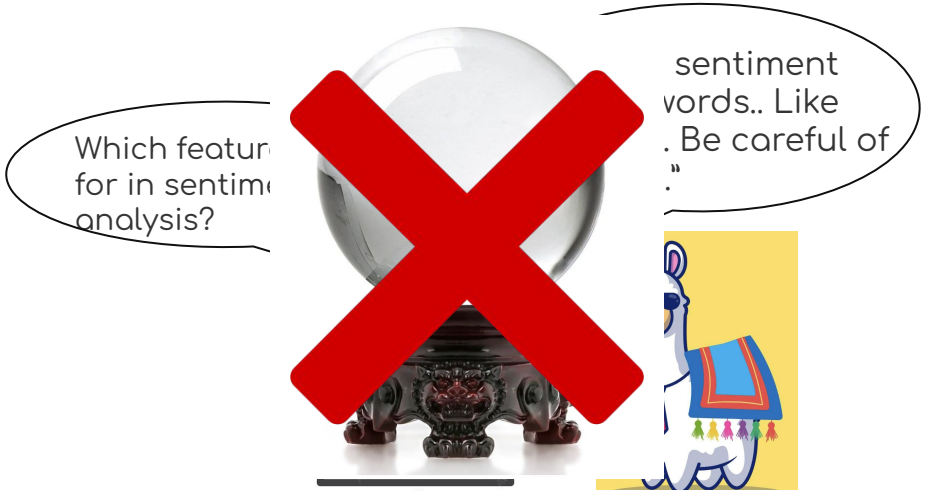
Image source: Amazon

2 ~~Disentangled representation~~

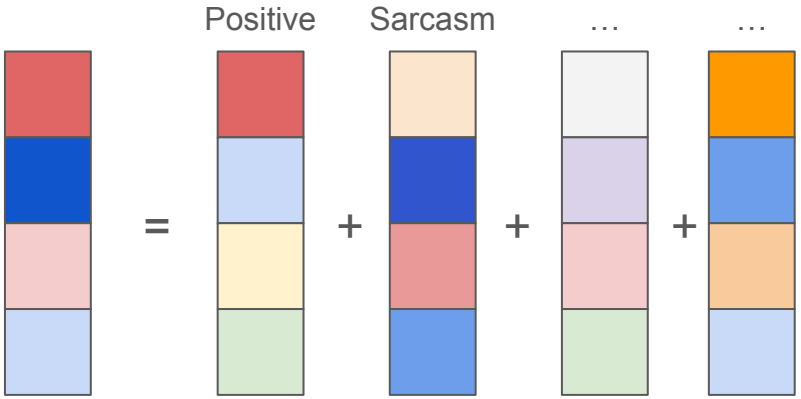


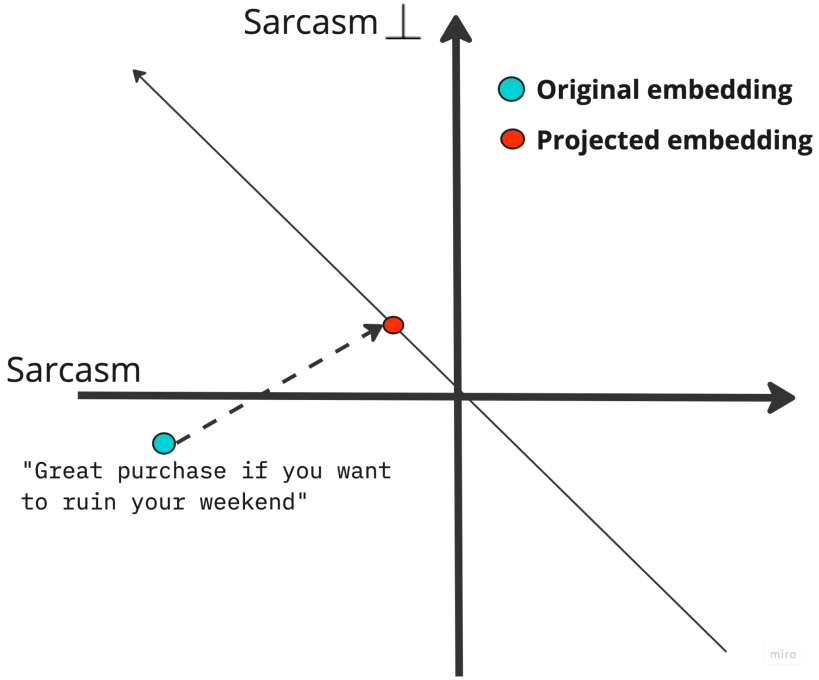
Is all hope lost?

Translating The Ideal World To Our Reality



Translating The Ideal World To Our Reality





LLMs as Oracles

You

I am trying to classify review sentiment as positive or negative. Which linguistic features are potentially misleading/spurious?

ChatGPT

Sarcasm and Irony:

- Sentiment expressed through sarcasm or irony can be challenging for models to detect accurately. Phrases may convey the opposite sentiment from what is explicitly stated.

Harmful concepts insights

Helpful concepts insights

You

I am trying to classify review sentiment as positive or negative. Which linguistic features should i look for?

ChatGPT

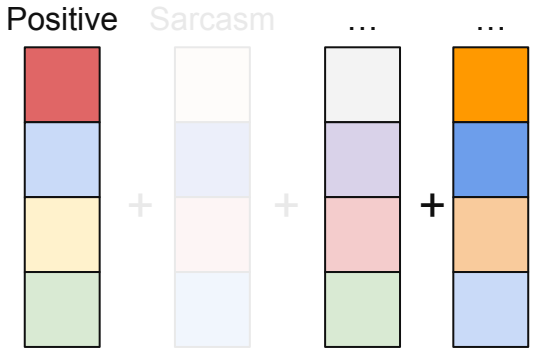
Sentiment Words:

- Look for sentiment-bearing words that explicitly convey positive or negative emotions (e.g., "happy," "satisfied" for positive, and "unhappy," "unsatisfied" for negative).

Embedding Debiasing as Proxy to Disentangled Representation

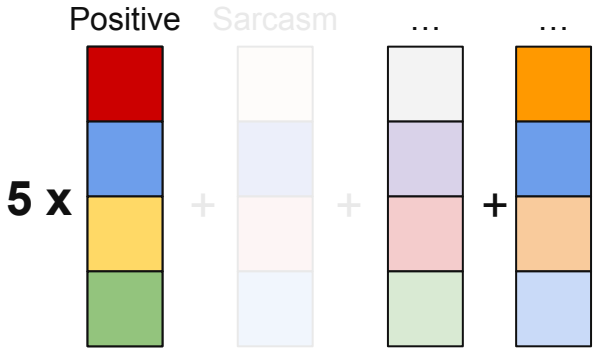
Neutralize harmful components

$$\hat{x} \leftarrow x - \frac{\langle x, v^{\text{harmful}} \rangle}{\langle v^{\text{harmful}}, v^{\text{harmful}} \rangle} v^{\text{harmful}}$$



Enhance useful features

$$\hat{x} \leftarrow \hat{x} + \frac{\langle \hat{x}, v^{\text{helpful}} \rangle}{\langle v^{\text{helpful}}, v^{\text{helpful}} \rangle} v^{\text{helpful}}$$



RoboShot

List the biased/spurious differences between waterbirds and landbirds

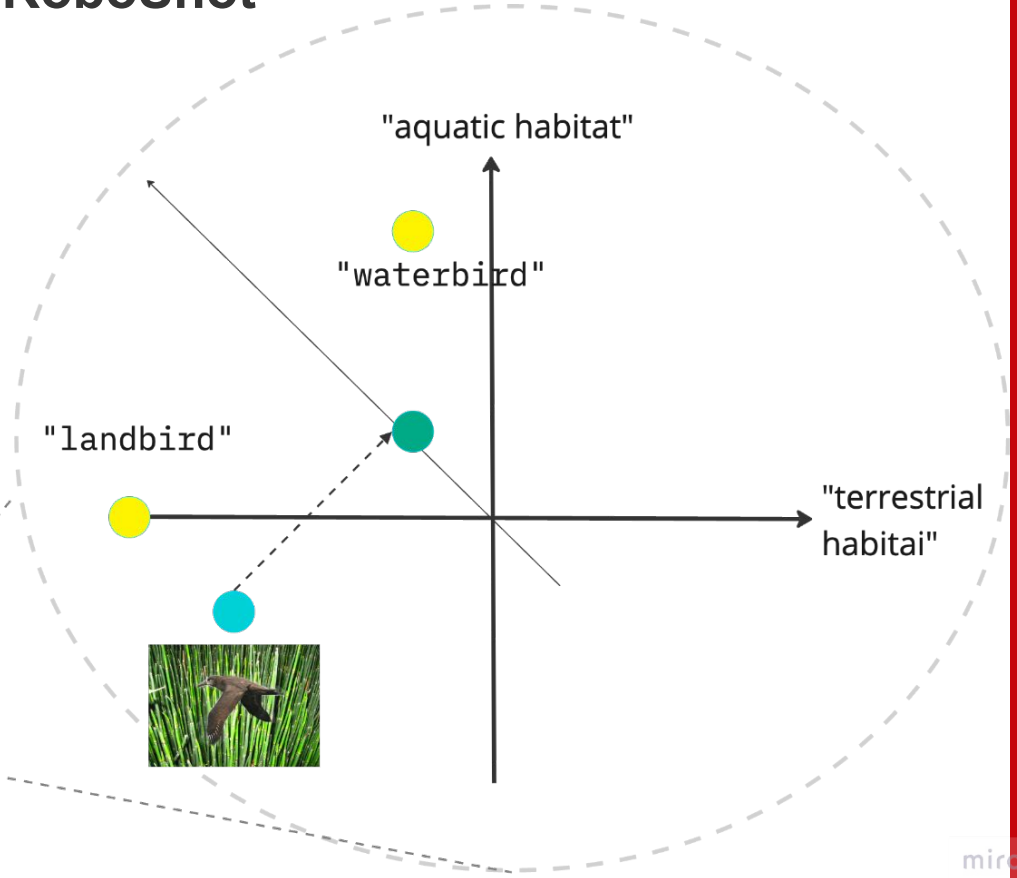
LLM

['aquatic habitat', 'terrestrial habitat']

Waterbird with land background

CLIP

["waterbird", "landbird"]



Future Work

1. Optimize the components
    a. Oracle: optimize prompts, use RAG
    b. Integrating oracle knowledge: intervene decoding layer, more sophisticated debiasing technique
2. Can we fine-tune models without labels?

Check out our paper!