Chalmers University of Technology
Unversity of Gothenburg
WASP—HS

Can Large Language Models (or Humans) Disentangle Text?

Nicolas Audinet de Pieuchon[1], Adel Daoud[2], Connor Thomas Jerzak[3], Moa Johansson[1], Richard Johansson[1]

[1] Chalmers University of Technology Göteborg [2] Institute for Analytical Sociology Linköping [3] University of Texas at Austin

Motivation

Disentanglement is the task of removing a forbidden variable Z from text while preserving as much of the text as possible

great little camera case . has a main pocket for the camera and a second smaller pocket to hold batteries and memory stick . it 's well made ballistic nylon - very durable . attaches to belt or sholder strap

Z = sentiment = positive
topic = camera

$\rightarrow d(\cdot) \rightarrow$

The camera case features a main pocket for the camera and a secondary pocket for accessories such as batteries and memory sticks. It is constructed from ballistic nylon. This case can be attached to a belt or shoulder strap

Z = sentiment = ?
topic = camera

Can be done at the text embedding level[1], but:

- Requires large set of annotated examples
- Less interpretable (no disentangled text)

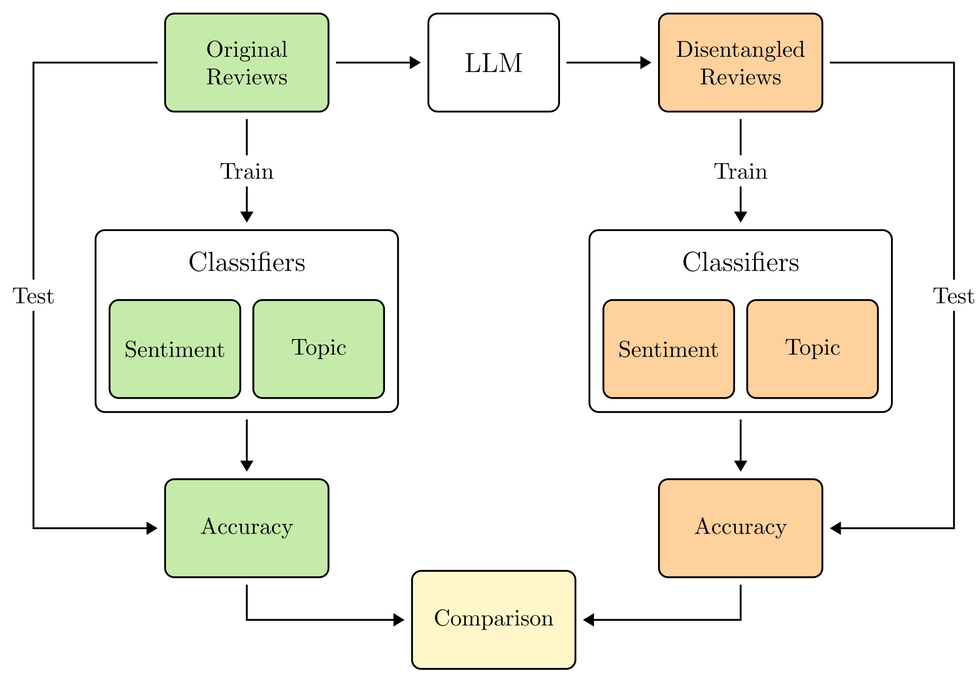
Can LLMs disentangle text out-of-the-box?

Can LLMs outperform humans at disentanglement?

Experiment Setup

Amazon reviews dataset[2]

- 2000 samples
- Two labels per sample: sentiment and topic
- Approximately balanced classes



Classifiers: logistic regression over DistilBERT embeddings

Prompting Strategies

Few-shot:

> Rewrite the review such that the sentiment is completely neutral. It is very important that one cannot tell whether the review is positive or negative at all. Try and keep all other information in the review.
>
> Here are a few examples of how to do this.
>
> Example 1: [...]
> Example 2: [...]
> Example 3: [...]
>
> Here's the review: [Review here]

Prompt chaining:

1) List parts of the review associated with the forbidden variable (few-shot)
2) Rewrite the review from stage 1 such that all traces of the forbidden variable are removed

Human experiment also used prompt chaining

Results

Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No disentanglement		0.885 ± 0.035	0.946 ± 0.026
Mean projection[1]		0.524 ± 0.054	0.946 ± 0.026
Human*	Prompt chaining	0.800 ± 0.145	0.842 ± 0.165
Mistral 7B	Paraphrase	0.891 ± 0.037	0.951 ± 0.024
	Few-shot	0.877 ± 0.023	0.951 ± 0.015
	Prompt chaining	0.841 ± 0.039	0.953 ± 0.023
GPT-4	Paraphrase	0.899 ± 0.034	0.951 ± 0.024
	Few-shot	0.824 ± 0.045	**0.955 ± 0.024**
	Prompt chaining	**0.757 ± 0.044**	0.945 ± 0.023

* Human distillation only tested on 152 reviews

1) The LLMs were unable to disentangle sentiment
2) GPT4 outperformed humans
3) Embedding methods worked well

Future work: how does variable separability affect disentanglement?

Check out the paper!

1 Pantea Haghighatkhah et al. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection.
2 John Blitzer et al. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.