

Can Large Language Models (or Humans) Disentangle Text?

**Nicolas Audinet de Pieuchon¹, Adel Daoud^{1,2,5},
Connor T. Jerzak³, Moa Johansson¹, Richard Johansson^{1,4}**

¹Chalmers University of Technology, Sweden

²Linköping University, Sweden

³University of Texas at Austin, USA

⁴University of Gothenburg, Sweden

⁵The Center for Advanced Studies in Behavioral Sciences, Stanford University, USA

NLP + CSS Workshop @NAACL, 2024

- **Disentangling Variables from Text:**

- Sometimes the text is ‘contaminated’ by a variable we want to remove or control for.
- E.g., removing demographic markers, sentiment, or personal info for fairness, causal inference, or ethical reasons.

- **Common Approaches:**

- Representation-level interventions (e.g., adversarial training, linear projections).
- BUT, these often require labeled data, yield transformations that are *not* easily human-interpretable.

- **Our Questions:**

1. *Can Large Language Models (LLMs) directly rewrite text to remove a target variable while preserving the rest?*
2. *How do humans perform on this same task?*

Defining Text Disentanglement

- **Representation Disentanglement (Prior Work):**
 - Guarding function $h(X) \perp Z$ for text representation X and variable Z .
 - E.g., removing gender info from a text embedding.
- **Our Extension to *Text-Level*:**
 - Transform raw text $W \mapsto d(W)$, so that $d(W) \perp Z$.
 - But preserve other important variables in W .
 - Minimally intrusive: we don't want to destroy all content.
- **High-Level Goal:**

$d(W) \perp Z$ while maximizing similarity to original text for other signals.

- **Target Variable (Z): Sentiment**

- Amazon Reviews dataset: 2000 reviews with binary sentiment (pos/neg) and topic/category.
- Chosen because sentiment typically spreads throughout the review, making this *a hard test*

- **LLM Approaches:**

- **Models:** Mistral 7B, GPT-4.
- **Prompt Strategies:**
 - 1 Paraphrase (control)
 - 2 Few-shot examples
 - 3 Prompt chaining (2-stage: identify + remove sentiment)

- **Comparisons:**

- **Representation projection (mean projection).**
- **Humans** performing manual text editing to remove sentiment.

Experimental Setup

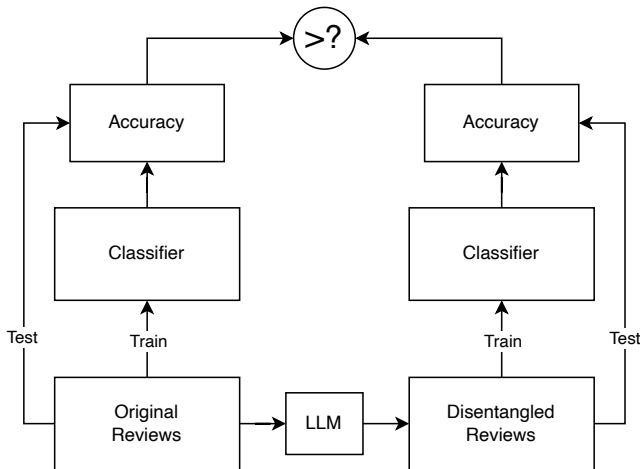


Figure: *Classifier approach for measuring disentanglement: train/test on original text vs. processed text for both sentiment and topic.*

- **Did we succeed in removing target signal?**
 - Assess accuracy of logistic regression on disentangled text embeddings.
 - If $d(W)$ truly removes sentiment, a classifier trained/tested on $d(W)$ *should* be near chance ($\sim 50\%$)
- **Did we preserve other text content?**
 - If $d(W)$ is *minimally intrusive*, a topic classifier should perform about as well as on original text.

Results

- **Mean Projection (Representation Level):**

- Achieves near chance for sentiment (about 52%), while preserving topic classification.
- But not interpretable at the text level.

- **LLMs:**

- Even GPT-4 cannot drive sentiment classifier below $\approx 76\%$ accuracy (still far above 50%).
- Topic remains well-preserved, near original performance.

- **Humans:**

- Also struggle; rewriting leaves *sentiment clues* for classifiers

Method	Sentiment Acc.	Topic Acc.
No Disentangling	88.5%	94.6%
Mean Projection	52.4%	94.6%
Human	80.0%	84.2%
Mistral 7B (best)	84.1%	95.3%
GPT-4 (best)	75.7%	94.5%

- **Limited Separability:**

- Both LLMs *and* humans often cannot fully strip sentiment from a text while retaining all else.
- Suggests sentiment is “baked in” throughout.

- **Challenges for Real-World Use:**

- If text can't be robustly disentangled, then representation-based methods might produce “incongruent” transformations unfaithful to real text.
- Raises caution about interpretability and fairness claims.

- **Next Directions:**

- More nuanced rewriting or advanced constraints.
- Larger curated datasets focusing on partial textual transformations (e.g., local vs. global manipulations).

Conclusion & Limitations

- **Conclusion:**

- Current LLMs do not reliably remove sentiment traces while preserving *all* other information.
- Even humans struggle, hinting that sentiment can be deeply entwined in text.

- **Limitations:**

- Some variables (e.g., localized personal info) might be easier to remove than sentiment.
- We evaluate with ML classifiers; real-world semantic cues differ.
- Further work needed for tasks where variables are heavily correlated (e.g., ideology and bias).

For more:

`nicaudinet.github.io`