Can Large Language Models (or Humans) Disentangle Text?

Citation for the original published paper (version of record):
Audinet De Pieuchon, N., Daoud, A., Jerzak, C. et al (2024). Can Large Language Models (or Humans) Disentangle Text?. Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024): 57-67.
http://dx.doi.org/10.18653/v1/2024.nlpcss-1.5

Can Large Language Models (or Humans) Disentangle Text?

Nicolas Audinet de Pieuchon[1] **Adel Daoud**[1,2,5] **Connor T. Jerzak**[3]
Moa Johansson[1] **Richard Johansson**[1,4]
[1]Chalmers University of Technology, Sweden [2]Linköping University, Sweden
[3]University of Texas at Austin, USA [4]University of Gothenburg, Sweden
[5]The Center for Advanced Studies in Behavioral Sciences, Stanford University, USA.
{nicolas.audinet, daoud, moa.johansson, richajo}@chalmers.se,
connor.jerzak@austin.utexas.edu

Abstract

We investigate the potential of large language models (LLMs) to disentangle text variables—to remove the textual traces of an undesired forbidden variable in a task sometimes known as text distillation and closely related to the fairness in AI and causal inference literature. We employ a range of various LLM approaches in an attempt to disentangle text by identifying and removing information about a target variable while preserving other relevant signals. We show that in the strong test of removing sentiment, the statistical association between the processed text and sentiment is still detectable to machine learning classifiers post-LLM-disentanglement. Furthermore, we find that human annotators also struggle to disentangle sentiment while preserving other semantic content. This suggests there may be limited separability between concept variables in some text contexts, highlighting limitations of methods relying on text-level transformations and also raising questions about the robustness of disentanglement methods that achieve statistical independence in representation space.

1 Introduction

When computational social scientists analyze text data there are situations where the text is contaminated by a *forbidden variable* that we want to preclude from our analysis or handle in a special way. For instance, methods in causal inference that use text to correct for unseen confounders (Keith et al., 2020; Roberts et al., 2020) run into difficulties if the text is influenced by the treatment variable (Daoud et al., 2022; Gui and Veitch, 2023). In other contexts, due to ethical, legal, or robustness considerations, we may want to ensure that models trained on a corpus are not influenced by effects such as demographic factors (Bolukbasi et al., 2016), domains (Ganin and Lempitsky, 2015), personal information (Li et al., 2018), or other sensitivity information (Hovy and Prabhumoye, 2021).

There are several methods that process text representations to enforce invariance with respect to a forbidden variable (Barrett et al., 2019; He et al., 2020; Ravfogel et al., 2020; Haghighatkhah et al., 2022; Belrose et al., 2023). With a few exceptions, these methods operate on a numerical *representation* of the text, and not directly on the text itself, which makes them less interpretable. Furthermore, they typically require a large set of annotated examples of the forbidden variable, which may not always be available or may be costly to obtain.

Computational social science using text as data is currently being transformed by the introduction of large language models (LLMs) applied in a zero-shot or few-shot fashion (Ziems et al., 2024; Törnberg, 2024). In this paper, we consider the question of whether LLMs have an out-of-the-box ability to *disentangle* a text: to transform it so that the value of a forbidden variable is hidden while preserving as much as possible of the original text. We apply the LLMs in a few-shot setup to reproduce the scenario where we have no large-scale annotations available of the forbidden variable. We investigate a variety of prompt-based techniques. If successful, the disentanglement process would also be understandable to a human reader, since the changes are carried out via interpretable transformations of the text itself.

We find that this type of disentanglement is challenging for the current generation of LLMs, as well as human annotators, to carry out. While the most powerful LLMs (e.g., GPT-4-class models) sometimes transform text so that it is difficult for humans to determine the original value of the forbidden variable, we see only a slight reduction in the accuracy of classifiers predicting the forbidden variable; its statistical presence is clearly maintained.

Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 57–67
June 21, 2024

2 Related work

Our work builds on a growing body of literature on removing undesired information from text representations. This line of research has focused on developing algorithms to learn representations that are independent of protected attributes like demographic variables (Li et al., 2018; Raff and Sylvester, 2018; Barrett et al., 2019; Belrose et al., 2023). These methods typically leverage adversarial training or projection techniques to encourage models to learn representations orthogonal to the forbidden variable.

In addition, some work has also used LLMs for manipulating text, sometimes with goals related to disentanglement in mind. For example, work has used language models for style transfer and controlled text generation, aiming to modify attributes like sentiment or formality while preserving core content (Mir et al., 2019; Malmi et al., 2020). Other research has leveraged language models for tasks like paraphrasing (Krishna et al., 2020), simplification (Martin et al., 2020), and neutralizing biased language through LLM fine-tuning (Ghanbarzadeh et al., 2023). We build on these works by exploring the use of out-of-the-box language models for the task of forbidden variable removal via text disentanglement. To the best of our knowledge, we are the first to study the effectiveness of LLM prompting for the disentanglement task systematically and to compare this approach to human performance.

3 Defining Disentanglement

Previous work has focused on removing a forbidden variable from a numerical text representation, not from the raw text itself. For a representation X and a forbidden variable Z, Ravfogel et al. (2023) defined *concept erasure* as a process that finds a *guarding function* h such that $h(X) \perp Z$. Several approaches have been proposed to find guarding functions; most work has focused on finding projections that optimize guardedness with respect to linear classifiers (Ravfogel et al., 2020; Belrose et al., 2023). However, this process is not interpretable in that a human cannot easily reason about what the transformation is doing in the numerical representation space. In addition, finding the guarding function requires a collection of annotated training instances.

In this article, we use a similar conceptualization, but we transform the text directly instead of working in a latent representation space. We define *text disentanglement* with respect to a forbidden variable Z as finding a guarding function d that takes a text and returns a transformed text where textual traces of Z have been removed so that $d(W) \perp Z$.

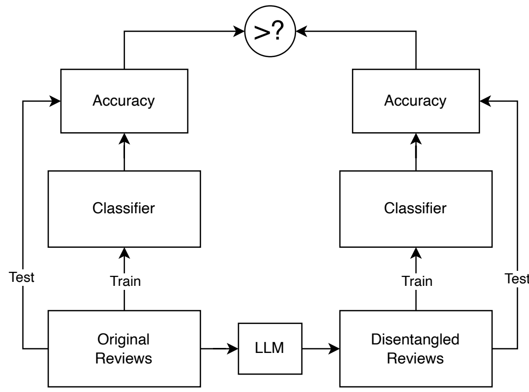

Figure 1: The experimental setup for measuring the effectiveness of LLMs at removing a target variable from the raw text representation.

We can trivially satisfy this independence criterion by letting d return a transformed text that is unrelated to the original text W (e.g., an empty text). To avoid such uninformative cases, we also want the transformation to be minimally intrusive. In previous work that relied on linear projections, this condition was satisfied by construction.

Although the meaning of intrusiveness will vary with research context, we here define it as our ability to measure the effect of disentanglement on our ability to predict other variables represented in the text. Looking ahead, we envision that minimal intrusiveness can be defined on a semantic level in terms of maximizing similarity in a representation space or on a superficial level by minimizing string edit distance; at this point, we leave this formalization to future work.

4 Method

The goal of our experiments is to determine whether current out-of-the-box LLMs are able to disentangle a forbidden variable from text without removing traces of other variables. We test two LLMs: Mistral 7B and GPT-4. We choose Mistral 7B as a smaller open-source LLM that performs well on benchmarks and can be run on consumer hardware (Jiang et al., 2023). We choose GPT-4 as a top-performing commercially available model (as of 2024) (Achiam et al., 2023). More specifically, we use Mistral 7B v0.2 (instruction-tuned) and "gpt-4-0125-preview" from the OpenAI API.

To add context to our results, we also test the performance of mean projection (Haghighatkhah et al., 2022), a method for removing specific information from representations, as well as, critically, the performance of human annotators. We include human annotators in our experiments to provide a benchmark for the level of forbidden variable removal that can be achieved by careful manual text editing, and to explore the intrinsic challenges that may stem from the entanglement of the forbidden variable and text content in natural language.

4.1 Dataset

In the experiments, we use a dataset consisting of 2000 Amazon reviews. This is a subset of the dataset published by Blitzer et al. (2007), which was originally created to investigate domain shifts and domain adaptation of sentiment classifiers. In our subset, each review has a label for the sentiment (positive or negative) and a label for the topic or product category (book, music, camera, health, DVD, or software). The dataset is approximately balanced in both sentiment and topic labels (e.g., 50.65% of reviews have a positive sentiment; the topic category with the largest number of reviews is "camera" with 17.8%; the category with the smallest number is "software" with 15.15%). We choose the sentiment as the forbidden variable and the topic as a proxy for the traces of other variables in the text we wish to keep.

We used this data because, in these reviews, sentiment information tends to be spread throughout the text, rather than localized to a specific sentence or section. This makes sentiment a challenging variable to remove while preserving other text content. As a result, this corpus serves as a strong test case—if LLMs (or humans) can successfully disentangle sentiment from these reviews while retaining the topic information—we would have compelling evidence of their text disentanglement capabilities. In other words, this test is strong in that we expect sentiment information to be spread throughout the text, as opposed to localized in a specific portion of the text, rendering the disentanglement task here more difficult than in the localized case.

4.2 Prompting

In order to instruct LLMs to disentangle a forbidden variable from text, we write the task description as a prompt. Given that the sentiment was chosen as the forbidden variable, the LLM is explicitly told to remove the sentiment from the reviews by revising them to be neutral. The LLM is also instructed

```
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
```

Figure 2: Excerpt from the few-shot prompt template. In our tests, $Review is replaced with the original text of each review.

to keep all other information when revising the reviews. The LLM is not explicitly instructed to keep information about topic, as we want topic to be a proxy for all non-sentiment semantic information.

We investigate two prompting strategies for constructing prompts: few-shot prompting and prompt chaining. Few-shot prompting involves giving one or more examples of how to solve the task as part of the prompt. Doing this has been shown to improve model performance over a zero-shot setup and is one of the most common prompting techniques (Brown et al., 2020). Figure 2 shows an excerpt of the few-shot prompt used in our experiments. The full prompt is shown in Figure 7 in Appendix A.2. We provide 3 task examples to the LLM.

We also investigate prompt chaining as an alternative prompting strategy (Wu et al., 2022). Prompt chaining involves breaking the problem into smaller tasks and asking the LLM to complete the tasks one at a time, keeping the prompts and answers from previous tasks in context. This strategy can boost performance and gives access to intermediate reasoning steps, which can be used to better understand how the model reasoned about the task.

For our purposes, we use a 2-stage prompt chain. In stage 1, the LLM is asked to return a list of the parts of the text that seem to be associated with the forbidden variable. We include 3 examples of how to solve the task as part of the prompt for the first stage. In stage 2, the LLM is asked to rewrite the review from the first stage such that all traces of the forbidden variable are removed. Like before, the LLM is also instructed to keep parts of the review not associated with the forbidden variable. The full prompts for both stages are available in Figure 8 and Figure 9 Appendix A.2.

To investigate the default rewriting behavior of the LLMs we also test a paraphrase prompt, which

Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No disentanglement		0.885 ± 0.035	0.946 ± 0.026
Mean projection		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

Table 1: The impact of disentangling text on sentiment and topic classification accuracy. Results are computed over 2000 Amazon reviews, except for the human setting which was computed on 152 reviews. ↑ and ↓ indicate whether higher values are or are not preferred.

asks the model to paraphrase the review without changing the meaning (see Figure 6 in Appendix A.2). If the LLM perfectly paraphrases the text without altering its semantic content, we would expect the rewritten text to exhibit the same level of association with the forbidden variable as the original text. However, if the LLM introduces any changes or artifacts during the rewriting process, this could amplify or diminish the signal of the forbidden variable compared to the initial text.

Finally, we also compare performance with a human baseline. In this baseline, we instructed three people via prompt chaining to (a) list out portions of the text related to sentiment and (b) re-write the text removing those sentiment-related portions but retaining all other information.

We make the raw reviews, the human-coded sentiment-related content, and the human-disentangled text available at `doi.org/10.7910/DVN/TEC1ZP` with replication repository at `GitHub.com/AIandGlobalDevelopmentLab/TextDisentanglement`.

4.3 Evaluation Design

We use two classifiers to evaluate the effectiveness of LLMs at disentanglement. Both classifiers are trained to label the sentiment of the reviews as this was our chosen forbidden variable. The first classifier is trained and tested on the original reviews; the second is trained and tested on the processed reviews. We then compare the accuracy of the two. Given that our dataset is approximately balanced, if the traces of the forbidden variable were successfully removed, we would expect the second classifier to have coin-toss accuracy. If traces of the forbidden variable were not removed, we would expect the two classifiers to have similar accuracy.

We use the same setup to test whether information about the topic is kept in the processed reviews. See Figure 1 for a summary of our evaluation setup.

We use logistic regression trained on document embeddings for all classifiers. The document embeddings are generated by taking the mean of all token embeddings within the document. Token embeddings are generated with DistilBERT (Sanh et al., 2019). To generate confidence intervals we bootstrap over the logistic regressions, using 500 bootstrap samples and a confidence level of 95%; we use an 80/20 train/test split.

5 Results

Our results are summarized in Table 1. In general, they show that current out-of-the-box LLMs are unable to consistently remove sentiment from data such as Amazon reviews. Of the LLMs and prompting strategies tested, GPT-4 with prompt chaining achieves best (i.e., lowest) sentiment classifier accuracy with an average of 75.7% (see an example in Appendix A.1). While the human annotators also struggled with removing traces of the sentiment from the reviews, they achieve results comparable with GPT-4 in sentiment accuracy. The mean projection experiment shows that removing almost all traces of sentiment from the reviews while keeping traces of the other variables is possible when operating at the representation level.

The LLMs performed well at keeping information about topic. Furthermore, the paraphrase prompting strategy shows that they successfully keep information about sentiment when rewriting: in the case of GPT-4, the prompt even leads to a slight increase in sentiment classifier accuracy. This indicates that LLMs may amplify original text signals when paraphrasing.

6 Implications

While the LLMs generally struggle to remove statistical associations with the forbidden variable, we also find that human coders face a similar difficulty. This finding implies that there may be, at least for the task described here, limited separability in the text between the forbidden variable and the remainder of the text. This limited separability raises questions about the robustness of disentanglement methods operating on downstream representations as such methods may inadvertently generate representations incongruous with real text. Further research is therefore needed to develop techniques that can effectively disentangle the forbidden variable from the relevant semantic content in a way that respects the content of the original text.

Ethical Considerations

Our work explores the capabilities and limitations of large language models and human annotators in disentangling text variables, raising several important ethical considerations. For example, our findings highlight the challenges of completely removing traces of a target variable while preserving other semantic content in the raw text space. This underscores the importance of transparency and interpretability when applying this or other disentanglement methods, as residual signals may still be detectable even after processing or the processed text representations may be incongruent with the original text semantics. It is important to communicate these limitations to end-users and decision-makers.

Limitations

Our experiments focus on variables where the relevant information is spread throughout the text, such as sentiment in product reviews. However, in some applications, the forbidden variable may be more localized and separable, such as personal information like names or addresses (Hovy and Prabhumoye, 2021). In these cases, the disentanglement task may be easier, as the target information can be more precisely removed.

We also focus on sentiment and topic variables that are relatively independent in our dataset. However, in real-world scenarios, variables of interest may be more intrinsically interrelated, such as political ideology and slant in news articles. Disentangling inherently correlated variables while preserving salient information could be more challenging (Daoud et al., 2022).

Finally, our evaluation results rely on machine learning classifiers, which may not fully capture human perception of the removal of the forbidden variable. Classifiers detect statistical patterns but do not necessarily "read" text like humans do. There could be cases where classifiers detect residual signals that are not semantically meaningful to humans or where important nuances are lost that are more apparent to humans. To get a more complete picture of disentanglement effectiveness, future work should augment machine evaluations with human judgment experiments, such as having annotators guess the original target variable from the disentangled text.

Acknowledgements

This research was supported by the project *Countering Bias in AI Methods in the Social Sciences* under the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Adel Daoud, Connor Jerzak, and Richard Johansson. 2022. Conceptualizing treatment leakage in text-based causal inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5638–5645, Seattle, United States.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada.

Lin Gui and Victor Veitch. 2023. Causal estimation for text data with (apparent) overlap violations. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8395–8416, Abu Dhabi, United Arab Emirates.

Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 279–285, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, United States.

Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada.

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–22, New York, NY, USA. Association for Computing Machinery.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Appendix

A.1 disentanglement Example

i bought this item because my husband heard is the best internet security . now i regret buying one . since ive installed it , it slowed my computer , its very very hard to close window and worse when i disconnected from the internet i have a hard time logging off . everytime im on the internet , i get mad and upset . its not worth all the stress . i definitely dont recommend this product . now im trying to figure how i can return the product and go back to norton

Figure 3: Example of an original review from the Amazon dataset

The parts of the review that contain information about the sentiment are:

∗ my husband heard is the best internet security
∗ now i regret buying one
∗ it slowed my computer
∗ its very very hard to close window
∗ worse when i disconnected from the internet i have a hard time logging off
∗ everytime im on the internet, i get mad and upset
∗ its not worth all the stress
∗ i definitely dont recommend this product
∗ now im trying to figure how i can return the product

Figure 4: GPT-4 response for the first stage of prompt chaining with the review from Figure 3

I purchased this item after hearing it was considered a top choice for internet security. Since installing it, I have observed changes in my computer's performance and behavior when disconnecting from the internet, as well as the process required to log off. I am currently exploring options to return the product and considering switching back to Norton.

Figure 5: GPT-4 response for the second stage of prompt chaining with the review from Figure 3 and the first stage response from Figure 4

A.2 Prompts

Paraphrase the following text without changing the meaning:

$REVIEW

Figure 6: Prompt template for the "identity" experiments.

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: if the original review was:

i bought this album because i loved the title song . it 's such a great song , how bad can the rest of the album be , right ? well , the rest of the songs are just filler and are n't worth the money i paid for this . it 's either shameless bubblegum or oversentimentalized depressing tripe . kenny chesney is a popular artist and as a result he is in the cookie cutter category of the nashville music scene . he 's gotta pump out the albums so the record company can keep lining their pockets while the suckers out there keep buying this garbage to perpetuate more garbage coming out of that town . i 'll get down off my soapbox now . but country music really needs to get back to it 's roots and stop this pop nonsense . what country music really is and what it is considered to be by mainstream are two different things .

then the neutral rewrite might be:

I bought this album because of the title song. The rest of the album I didn't know as well. Kenny Chesney is a popular artist in the Nashville music scene. He makes many albums with his record company. Country music has been evolving from its roots to a more pop sound.

Example 2: if the original review was:

this is a very good shaver for the private area . however , the key to getting the best results is to trim the longer hairs with scissors or the largest guard first . this will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work . the foil will not be able to do its job if the hairs are too long . the only problem i had with the shaver was that it did not enable me to shave my back like it claimed . however , i use the ' mangroomer ' back shaver for this and it is perfect for you to shave off all your back hair easily with its elongated handle . it is a great product as well . therefore , i would have to say these two products coupled together seem to cover all the bases for men 's grooming on the body . i would highly recommend both of them for perfect manscaping results

then the neutral rewrite might be:

To use this shaver in the private area it is important to trim the longer hairs with scissors or the largest guard first. This will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work. The foil will not be able to do its job if the hairs are too long. The shaver might also not work well on the back. For this, there are other options such as the 'Mangroomer' back shaver which has an elongated handle that makes it easy to shave back hair.

Example 3: if the original review was:

i bought bead fantasies and bead fantasies ii at the same time after reading the positive reviews ; i wish i had looked at these books before buying . there are pretty motifs that i will incorporate into my beading projects but i find the small typed directions overly simplistic and the diagrams are too small . i 'm glad this is n't my first beading book or i would feel totally discouraged from trying any of these projects . i wo n't be buying bead fantasies iii . the art and elegance of beadweaving and coraling technique remain my favorite beading books .

then the neutral rewrite might be:

I bought Bead Fantasies and Bead Fantasies II at the same time. I like some of the motifs but not others. This is not my first beading book. The art and elegance of beadweaving and coraling technique are great beading books.

Here's the review:

$REVIEW

Figure 7: Prompt template for the "few-shot" experiments.

Identify the places in the following review which contain information about the sentiment and return them as bullet points.

Here are a few examples of how to do this.

Example 1: if the original review was:

i bought this album because i loved the title song . it 's such a great song , how bad can the rest of the album be , right ? well , the rest of the songs are just filler and are n't worth the money i paid for this . it 's either shameless bubblegum or oversentimentalized depressing tripe . kenny chesney is a popular artist and as a result he is in the cookie cutter category of the nashville music scene . he 's gotta pump out the albums so the record company can keep lining their pockets while the suckers out there keep buying this garbage to perpetuate more garbage coming out of that town . i 'll get down off my soapbox now . but country music really needs to get back to it 's roots and stop this pop nonsense . what country music really is and what it is considered to be by mainstream are two different things .

then the parts of the review that contain information about the sentiment are:

∗ i loved the title song
∗ it 's such a great song
∗ the rest of the songs are just filler and are n't worth the money
∗ it 's either shameless bubblegum or oversentimentalized depressing tripe
∗ the suckers out there keep buying this garbage to perpetuate more garbage
coming out of that town
∗ but country music really needs to get back to it 's roots
∗ nonsense

Example 2: if the original review was:

this is a very good shaver for the private area . however , the key to getting the best results is to trim the longer hairs with scissors or the largest guard first . this will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work . the foil will not be able to do its job if the hairs are too long . the only problem i had with the shaver was that it did not enable me to shave my back like it claimed . however , i use the ' mangroomer ' back shaver for this and it is perfect for you to shave off all your back hair easily with its elongated handle . it is a great product as well . therefore , i would have to say these two products coupled together seem to cover all the bases for men 's grooming on the body . i would highly recommend both of them for perfect manscaping results

then the parts of the review that contain information about the sentiment are:

∗ this is a very good shaver for the private area
∗ the only problem i had with the shaver was that it did not enable me to shave
my back like it claimed
∗ it is perfect for you to shave off all your back hair easily with its
elongated handle
∗ it is a great product as well
∗ i would highly recommend

Example 3: if the original review was:

i bought bead fantasies and bead fantasies ii at the same time after reading the positive reviews ; i wish i had looked at these books before buying . there are pretty motifs that i will incorporate into my beading projects but i find the small typed directions overly simplistic and the diagrams are too small . i 'm glad this is n't my first beading book or i would feel totally discouraged from trying any of these projects . i wo n't be buying bead fantasies iii . the art and elegance of beadweaving and coraling technique remain my favorite beading books .

then the parts of the review that contain information about the sentiment are:

∗ i wish i had looked at these books before buying
∗ there are pretty motifs
∗ i find the small typed directions overly simplistic
∗ the diagrams are too small
∗ i 'm glad this is n't my first beading book
∗ i would feel totally discouraged
∗ i wo n't be buying bead fantasies iii

Here is the review:

$REVIEW

Figure 8: Prompt template for the first stage of the “prompt chaining” experiments.

Rewrite the original review such that all the information identified about the sentiment is removed. The goal is to make the review completely neutral. It is very important that one cannot tell whether the review is positive or negative at all. Keep all other information in the review.

Figure 9: Prompt template for the second stage of the “prompt chaining” experiments.