

NeuroCounterfactuals: Beyond Minimal-Edit Counterfactuals for Richer Data Augmentation

Phillip Howard[◇] Gadi Singer[◇] Vasudev Lal[◇] Yejin Choi^{♡♣} Swabha Swayamdipta^{♣♠}
[◇]Intel Labs [♣]Allen Institute for AI [♠]University of Southern California
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
phillip.r.howard@intel.com

Abstract

While counterfactual data augmentation offers a promising step towards robust generalization in natural language processing, producing a set of counterfactuals that offer valuable inductive bias for models remains a challenge. Most existing approaches for producing counterfactuals, manual or automated, rely on small perturbations via minimal edits, resulting in simplistic changes. We introduce NeuroCounterfactuals, designed as loose counterfactuals, allowing for larger edits which result in naturalistic generations containing linguistic diversity, while still bearing similarity to the original document. Our novel generative approach bridges the benefits of constrained decoding, with those of language model adaptation for sentiment steering. Training data augmentation with our generations results in both in-domain and out-of-domain improvements for sentiment classification, outperforming even manually curated counterfactuals, under select settings. We further present detailed analyses to show the advantages of NeuroCounterfactuals over approaches involving simple, minimal edits.

1 Introduction

Despite the enormous successes in natural language processing, out-of-domain (OOD) generalization still poses a challenge for even the most powerful models, which achieve remarkable performance in domain (Recht et al., 2019; Torralba and Efros, 2011). This can be attributed to the models’ reliance on spurious biases (Geirhos et al., 2020; McCoy et al., 2019; Gururangan et al., 2018), i.e. features which co-occur with the ground truth without any causal dependence (Simon, 1954). Adopting methods from causal inference (Pearl, 2009; Feder et al., 2022), training data augmentation with counterfactuals (CFs) has been proposed for NLP as one potential solution (Levesque et al., 2012; Kaushik et al., 2019, 2021). Counterfactuals are designed to study the change in a response variable (e.g., the

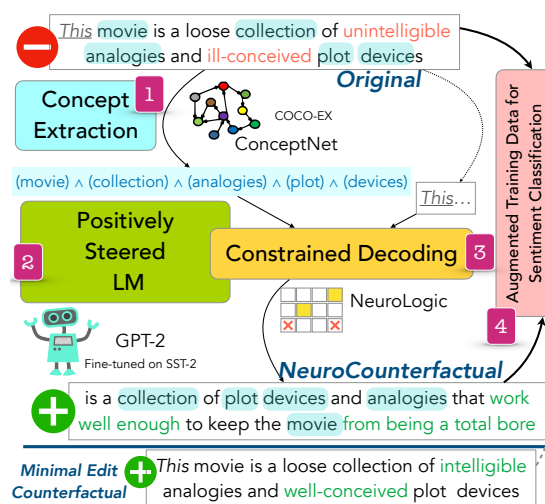


Figure 1: Illustration of our approach. **1** We extract tokens from an Original (negative) movie review that evoke concepts from ConceptNet (§2.1). **2** We use a GPT-2 model adapted to only reviews with the opposite (positive) polarity as a sentiment steer (§2.2). **3** Finally, to ensure that the generation is similar to the original, we use NeuroLogic, a constrained decoding approach (§2.3; Lu et al., 2021), where the constraints are extracted tokens from **1**. This results in NeuroCounterfactuals, which are loose counterfactuals of the original, but are more naturalistic (§4; Tab. 1), compared to minimal edit counterfactuals (bottom). **4** When used to augment training data for sentiment classification, our generations are valuable for OOD generalization (§3).

target label), following an intervention (e.g., altering a causal feature), typically in the form of edits to the input text (Khashabi et al., 2020; Andreas, 2020). Training data augmentation with counterfactuals can thus provide strong inductive biases to help with robustness against spurious biases, resulting in improved OOD generalization (Vig et al., 2020; Eisenstein, 2022).

However, designing the appropriate interventions to produce counterfactuals can be challenging. Indeed, most counterfactuals are produced via basic edits to the input text, either manually (Gardner

et al., 2020; Kaushik et al., 2019) or automatically (Yang et al., 2021; Wang and Culotta, 2021; Wu et al., 2021), such that the target label changes. These minimal edits are made via substitution, insertion or deletion of tokens in the original sentence, resulting in simplistic generations, which are often unrealistic and lack linguistic diversity.¹ As a result, counterfactuals via minimum edits often fail to provide adequate inductive biases to promote robustness (Khashabi et al., 2020; Huang et al., 2020; Joshi and He, 2022).

In this paper, we investigate the potential of more realistic and creative counterfactuals, which go beyond simple token-level edits, towards improving robust generalization. While allowing larger edits reduces proximity to the original sentence, we believe that this is a worthwhile trade-off for more realistic and creative counterfactuals, which offer greater flexibility in sentiment steering, increasing the likelihood that the counterfactual possesses the desired label. We propose a novel approach that can generate diverse counterfactuals via concept-controlled text generation, illustrated in Figure 1. In particular, our approach combines the benefits of domain adaptive pretraining (Gururangan et al., 2020) for soft steering of the target label (Liu et al., 2021), with those of NeuroLogic decoding (Lu et al., 2021), an unsupervised, inference-time algorithm that generates fluent text while strictly satisfying complex lexical constraints. As constraints, we use tokens that evoke salient concepts derived from ConceptNet (Speer et al., 2017). Our resulting generations, called NeuroCounterfactuals², provide loose counterfactuals to the original, while demonstrating nuanced linguistic alterations to change the target label (§2).

Compared to minimal-edit counterfactuals, our counterfactuals are more natural and linguistically diverse, resulting in syntactic, semantic and pragmatic changes which alter the label while preserving relevance to the original concepts (Table 1). On experiments with training data augmentation for sentiment classification, our approach achieves better performance compared to competitive baselines using minimal edit counterfactuals (§3). Our performance even matches baselines using human-annotated counterfactuals, on some settings, while avoiding the cost of human annotation. While Neu-

roCFs are designed to be loose counterfactuals, our detailed analyses show that it is still important to augment training data with examples possessing a moderately high degree of similarity with the original examples (§4). When the ultimate goal is improving robust generalization, we show that going beyond minimal edit counterfactuals can result in richer data augmentation.³

2 NeuroCounterfactuals

We describe our methodology for automatic generation of loose counterfactuals, NeuroCFs, for sentiment classification. The key idea underlying our approach is the need for retention of concepts to ensure content similarity to the original text, while steering the sentiment to the opposite polarity. Our method, illustrated in Figure 1, combines a concept-constrained decoding strategy with a sentiment-steered language model. First, we detail our approach for extracting the salient concepts from a document (§2.1). Next, we discuss language model adaptation to produce sentiment-steered LMs (§2.2). Finally, we provide an overview of the NeuroLogic decoding algorithm for controlled text generation, and how it can be adapted for the task of generating sentiment counterfactuals (§2.3).

2.1 Extracting Salient Concepts

Our first step constitutes extraction of concepts from the original document, which can be used to reconstruct its content, when used as constraints during decoding (§2.3). Specifically, we aim to identify a set of constraints which will require the counterfactual to be similar in content to the original sentence while still allowing the generation to be steered towards the opposite polarity. Using extracted concepts as constraints achieves this because the concepts consist of the content-bearing noun phrases as opposed to the sentiment-bearing adjectives. For example, in the original sentence from Figure 1, we seek to constrain our generated counterfactual to contain concept-oriented phrases, such as “*movie*”, “*analogy*”, and “*plot devices*” without explicitly requiring the presence of other tokens which may indicate the sentiment (e.g., “*unintelligible*”, “*ill-conceived*”).

We achieve this mapping via linking tokens and phrases in the document to nodes in the ConceptNet knowledge graph (Speer et al., 2017), thus evoking

¹For instance, the minimal edit counterfactual in Figure 1 contains the phrase “*loose collection of intelligible analogies*”, a somewhat unnatural construction for a positive movie review.

²NeuroCFs, for short.

³Our code and data are available at <https://github.com/IntelLabs/NeuroCounterfactuals>

Source	Label	Review
Original	⊖	But this <i>film</i> decided to throw away the <i>talents</i> of the <i>people</i> involved in a simpering <i>version</i> so watered down from the <i>source material</i> that it's amazing they had the <i>guts</i> to call it Wuthering <i>Heights</i> at all.
W&C.	⊖	But this film decided to throw away the talents of the people involved in a simpering version so watered down from the source material that it s unimpressive they had the guts to call it wuthering heights at all
Y.et al.	⊖	But this film decided to throw away the talents of the people involved in a simpering version so watered down from the source material that it's amazing they had the guts to call it wuthering heights at all.
NeuroCFs-1g	⊕	But the <i>film</i> <i>guts</i> its <i>source material</i> , and it does so with a version of the <i>heights</i> of artistry that <i>people</i> have come to expect from the <i>talents</i> of jean renoir.
NeuroCFs-np	⊕	But this <i>film</i> decided to take the <i>talents</i> of the <i>source material</i> and make them its own, and it's a <i>gutsier version</i> of the <i>people</i> we know and love from the <i>heights</i> .
Original	⊕	Unfortunately I had to rent a <i>Dreamcast</i> to play it, but even though I did beat it I can't wait to buy it for <i>PS2</i> .
W&C.	⊕	Fortunately i had to rent a dreamcast to play it but even though i did beat it i can t wait to buy it for ps2
Y.et al.	??	Unfortunately i had to rent a dreamcast to play it, but even though i did beat it i can't wait to buy it for ps2.
NeuroCFs-1g	⊖	Unfortunately it's not nearly as good as the <i>dreamcast ps2</i> version.
NeuroCFs-np	⊖	Unfortunately i had to rent a <i>dreamcast</i> to play it but even though i did beat it i can't recommend it for <i>ps2</i> or xbox.

Table 1: Comparison of IMDB-S train examples (Original) with generated counterfactuals from different approaches: W&C. (Wang and Culotta, 2021), Y.et al. (Yang et al., 2021), and our NeuroCF variants, designed to flip the target label. The sentiment labels for the counterfactuals can be ⊕ (positive), ⊖ (negative), or ?? (unclear), as assessed by authors of this work. For the baselines, substitutions and insertions are underlined, ignoring punctuation and capitalization, and ~~deletions~~ are struck out. NeuroCFs result in more complex changes to the original, and are more successful in steering the sentiment for label flipping; minimal edits are at times unable to result in meaningful changes to the sentiment, and result in reduced grammaticality. Concepts in the original sentence that were used as constraints to generate NeuroCFs are *in blue italics*. Also see App §A; Tab. 14.

salient concepts. Nodes in ConceptNet are represented as non-canonicalized, free-form text. To this end, we use COCO-EX (Becker et al., 2021), a ConceptNet entity linking tool. COCO-EX improves upon simple string-matching techniques which have been commonly used for ConceptNet entity linking in the past by selecting meaningful concepts and mapping them to a set of concept nodes based on relational information in the graph. Most extracted concepts correspond to nominal entities. Moreover, this mapping implicitly ensures that our extraction refrains from sentiment-bearing tokens and phrases.

We primarily use COCO-EX for its ability to identify meaningful concepts, but also explore the use of links to related concepts it provides in Section 4.4. We also compare with a baseline using noun chunks as constraints in App C.2.

2.2 Steering Sentiment via LM Adaptation

The second component for our method is a sentiment “steer”, i.e. an autoregressive language model which has been trained or adapted via finetuning (Gururangan et al., 2020) exclusively on sentences with single (negative or positive) polarity. Specifically, we use two steers for each sentiment label: one which models positive sentiment text, (denoted p_{θ}^{+}), and another which models negative sentiment text, (denoted p_{θ}^{-}), where θ indicates the parameters of the adapted language model. In contrast

to the hard predicate constraints over specific tokens as given by the extracted concepts in §2.1, our selective use of steering LMs can be viewed as a softer type of constraint which biases the generations towards text containing the desired sentiment polarity (Liu et al., 2021).

2.3 Decoding with Conceptual Constraints

Our method utilizes NeuroLogic Decoding (Lu et al., 2021), a controlled text generation algorithm to generate fluent text satisfying a set of lexical constraints from a pretrained language model. Given a series of predicates $D(\mathbf{a}, \mathbf{y})$ which are true iff \mathbf{a} appears in the generated sequence \mathbf{y} , NeuroLogic accepts a set of *clauses* $\{C_i \mid i \in 1, \dots, m\}$ consisting of one or more predicates specified in Conjunctive Normal Form (CNF):

$$\underbrace{(D_1 \vee D_2 \cdots \vee D_i)}_{C_1} \wedge \cdots \wedge \underbrace{(D_k \vee D_{k+1} \cdots \vee D_n)}_{C_m}$$

where each predicate D_i is a positive constraint, $D(\mathbf{a}_i, \mathbf{y})$, which is satisfied (i.e., evaluates as true) if the subsequence \mathbf{a}_i appears in the generated sequence \mathbf{y} .

NeuroLogic employs a beam search approximation of an objective function which maximizes the probability of the generated sequence while penalizing deviations from the set of m clauses:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} p_{\theta}(\mathbf{y}|\mathbf{x}) - \lambda \sum_{j=1}^m (1 - C_j) \quad (1)$$

where $\lambda \gg 0$ penalizes deviations from the set of constraints. Candidates are scored at each stage t of beam search according to their partial or full satisfaction of the constraints:

$$f(\mathbf{y}_{\leq t}) = \log p_{\theta}(\mathbf{y}_{\leq t}|\mathbf{x}) + \lambda \max_{D(\mathbf{a}, \mathbf{y}_{\leq t})} \frac{|\hat{\mathbf{a}}|}{|\mathbf{a}|} \quad (2)$$

where $\hat{\mathbf{a}}$ represents a subsequence of \mathbf{a} in the current generation and the maximum is taken over all unsatisfied constraints consisting of more than one token. This has the effect of preferring candidates which at least partially satisfy multi-token constraints; for example, a generated sequence $\mathbf{y}_{\leq t} = \text{“The boy climbs an apple”}$ would be rewarded for partially satisfying the constraint $\mathbf{a} = \text{“apple tree”}$ via its subsequence $\hat{\mathbf{a}} = \text{“apple”}$.

Unlike the top- k selection strategy used in traditional beam search, NeuroLogic performs pruning, grouping, and selection steps to identify the best candidates which satisfy the given constraints. Specifically, candidates which irreversibly violate one or more constraints are pruned, and the remaining candidates are grouped according to their number of satisfied clauses in order to encourage diversity. The best candidate within each group is then selected according to the scoring function in Equation 2.

Each word or phrase in the original example which is linked to a ConceptNet node (§2.1) becomes a clause in our constraint set used with NeuroLogic. We allow each clause to be satisfied by the lowercase or capitalized form of the concept via an OR constraint. For the example in Figure 1, this constraint set would be specified in CNF as follows:

$$(Movie \vee movie) \wedge (Plot Devices \vee plot devices) \wedge (Collection \vee collection) \wedge (Analogies \vee analogies)$$

Once the constraints have been identified in the original, we substitute the sentiment-steered LMs (§2.2) into Equation 1, corresponding to a polarity opposite to the original:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} p_{\theta}^i(\mathbf{y}|\mathbf{x}) - \lambda \sum_{j=1}^m (1 - C_j). \quad (3)$$

Here, $p_{\theta}^i = p_{\theta}^+$ when we aim to generate a positive-sentiment example and $p_{\theta}^i = p_{\theta}^-$, for a negative-sentiment example. The resulting generation, $\hat{\mathbf{y}}$, is a **NeuroCounterfactual (NeuroCF)**.

In Eq. 3, the generation is conditioned on \mathbf{x} , which indicates a prompt, comprising a prefix of

the original input; we investigate two variants for \mathbf{x} . When \mathbf{x} is a unigram ($1g$) comprising the *first token* of the original input, we call the generations **NeuroCFs-1g**. When \mathbf{x} is the *longest neutral prefix* of the original input, we call the generations **NeuroCFs-np**; these are slightly tighter NeuroCFs containing a greater portion of the original input. Table 1 provides examples showing the original sentence and our generated NeuroCFs, highlighting words in the original that were included in the concept-oriented constraint set for NeuroLogic decoding. NeuroCFs are not guaranteed to *not* contain new concepts, beyond the specifications of the constraint set. See App. §A for further examples.

3 Data Augmentation with NeuroCFs

Our experiments compare NeuroCFs to CFs from minimal edit approaches, for augmentation of sentiment classification training data.

3.1 Experimental Setup

Sentiment Steer Our positive and negative sentiment steers are based on a GPT-2 Large model (Radford et al., 2019), finetuned on (positive and negative, resp.) subsets of the Stanford Sentiment Treebank (SST-2; Socher et al., 2013) corpus, including train, test and validation splits.⁴

NeuroLogic For decoding with NeuroLogic, we use a beam size of 20, length penalty of 0.3, and an n -gram size of 2 for preventing repetitions. We use $\beta = 1.25$ as the reward factor for in-progress constraint satisfaction and set the constraint satisfaction tolerance to 2. Please refer to Lu et al. (2021) for details on these hyperparameters.

For the generation of NeuroCFs-np, we identify the longest neutral prefix of the original input. As candidates, we consider all prefixes containing at least 4 tokens, such that the rest of the review contains at least one identified concept. We filter the longest candidate, predicted as neutral using an off-the-shelf 5-way sentiment classifier.⁵

Following prior work (Kaushik et al., 2019), we generate NeuroCFs for a subset of movie reviews from the Internet Movie Database (IMDB; Pang and Lee, 2005), comprising 2440 examples randomly sampled and split into 70% training, 10% validation, and 20% test partitions, for a sentiment classification task (Maas et al., 2011). We augment the training data of a sentence-level version

⁴We use the [sentiment experts](#) released by Liu et al. (2021).

⁵From [ShannonAI](#).

of this dataset (**IMDB-S**)⁶, introduced by Wang and Culotta (2021); see App. B.1 for details.

3.2 Baselines: Other CF sources

We compare with multiple sentiment classification baselines employing counterfactuals for training data augmentation. Kaushik et al. (2019) crowdsource counterfactuals for IMDB, by soliciting minimal revisions to maintain coherence while flipping the sentiment, creating both a counterfactually augmented train as well as test dataset. We also consider two approaches that produce automatically generated counterfactuals. Wang and Culotta (2021) generate counterfactuals by automatically identifying causal words in the original example and substituting them with their antonyms, ensuring minimal edits. Similarly, Yang et al. (2021) automate counterfactual generation through the identification of causal terms which are either removed or replaced; they then filter candidates using MoverScore (Zhao et al., 2019) to ensure minimal edits were made to the original example. For all the baselines above, we train on sentence-level IMDB reviews, as well as sentence-level variants of the counterfactuals.⁷

Sentiment Classifier We compare several models, based on a RoBERTa-base architecture (Liu et al., 2019). Each model is trained on a counterfactually augmented dataset, where the CFs are either obtained via baselines above, or via our approaches (§2). We additionally train a baseline only on the original IMDB-S training data, without any CFs. Details on model training are provided in App. B.3.

Evaluation We report classification accuracy on a combination of in-domain and out-of-domain test sets. As in-domain test sets, we evaluate on the IMDB test set. We also evaluate on CFs for IMDB, crowdsourced by Kaushik et al. (2019). In addition, we evaluate on contrast sets (Gardner et al., 2020), which are expert-annotated CFs for IMDB test data. As another in-domain test set, we evaluate on the SST-2 movie reviews test set.⁸ Wu et al. (2021) produce task-agnostic, minimal edit counterfactuals with fine-grained semantic controls over different

types of perturbations, followed by human labeling; we also evaluate on these so-called Polyjuice CFs for SST-2 test.⁹ While SST-2 differs from IMDB in terms of word length and style, we nevertheless consider it in-domain for the purpose of our evaluations because both datasets are comprised of movie reviews.

For the OOD test sets, we consider the following binary sentiment classification datasets:

- The **Amazon** dataset (Ni et al., 2019) consists of consumer product reviews in the categories of software, fashion, appliances, beauty, magazines, and gift cards.
- The **Twitter** dataset (Rosenthal et al., 2017) from SemEval-2017 Task 4 contains social media posts collected from Twitter.
- The **Yelp** dataset¹⁰ contains consumer reviews originating from the Yelp dataset challenge.

3.3 NeuroCFs for Train Data Augmentation

Table 2 shows our results. NeuroCFs outperform alternative methods for automatic CF generation across every in-domain as well as OOD setting, including performance on CF test sets. The only exception is IMDB test, where we match the performance of the best approach (up to standard deviation). Across most CF and OOD test sets, the magnitude of our improvements is similar to or greater than the amount by which existing methods improve on the no-counterfactual baseline. Furthermore, most of these improvements are statistically significant ($p \leq 0.05$) relative to the results of both Yang et al. (2021) and Wang and Culotta (2021). NeuroCFs even surpass the performance of augmentation with crowdsourced counterfactuals from Kaushik et al. (2019) on most OOD settings. However, training on manual CFs results in higher performance when tested on human-written CFs; this might be attributed to distributional similarities (Geirhos et al., 2020; Koh et al., 2020). Regardless, our performance is close enough, despite using fewer training instances while avoiding the significant cost of human annotation.

Both NeuroCF-variants have comparable performance, with the NeuroCFs-np faring better on 4/8 benchmarks. Consistent with prior work (Wang and Culotta, 2021), we observe that training with CFs generally results in similar or slightly worse

⁶Initial experiments with NeuroLogic decoding with full length IMDB paragraphs were prohibitively slow, which we circumvented by using the sentence-level version.

⁷App. B.2 provides further details on our datasets.

⁸While our sentiment steers are trained on SST-2 data, we use NeuroLogic decoding to obtain counterfactuals for IMDB, which we use to train our sentiment classifier. Hence, it is unlikely that the classifier is exposed to SST-2 directly.

⁹We cannot compare with a baseline trained on Polyjuice CFs, as these are not available for IMDB, and would need human labeling.

¹⁰<https://www.yelp.com/dataset>



Source of CFs	$ D_{\text{train}} $	IMDB			SST-2		Out-of-domain		
		Test	CF (K. et al.)	Cont.Sets	Test	PolyJuice CFs	Twitter	Yelp	Amazon
None	8,173	93.22 _{0.42}	92.07 _{1.04}	86.85 _{1.06}	90.30 _{0.97}	84.74 _{0.46}	77.94 _{1.72}	94.71 _{0.67}	90.35 _{1.03}
Yang et al., 2021	10,376	92.15 _{0.79}	91.99 _{1.56}	86.67 _{1.46}	89.46 _{0.95}	86.90 _{0.57}	76.37 _{1.96}	94.23 _{0.59}	89.97 _{1.07}
Wang and Culotta, 2021	10,744	92.88 _{0.45}	94.03 _{0.91}	89.69 _{0.87}	89.26 _{1.55}	85.97 _{0.69}	77.09 _{1.97}	94.47 _{0.61}	90.88 _{0.89}
NeuroCFs-1g	15,437	92.60 _{0.59}	93.36 _{0.71}	89.04 _{1.02}	92.63 _{0.44}	87.11 _{0.52}	<u>77.98</u> _{1.22}	95.01 _{0.22}	92.32 _{0.51}
NeuroCFs-np	12,905	92.66 _{0.46}	95.03 _{0.47}	90.85 _{0.84}	<u>92.27</u> _{0.39}	88.35 _{0.41}	78.80 _{1.22}	94.51 _{0.87}	<u>92.24</u> _{0.71}
 Kaushik et al., 2019	16,679	92.63 _{0.48}	97.34 _{0.37}	95.22 _{0.45}	89.73 _{0.76}	90.10 _{0.29}	81.28 _{1.60}	93.94 _{0.52}	91.96 _{0.44}

Table 2: Sentiment classification accuracies, comparing IMDB-S training data augmentation with NeuroCFs vs. other sources of counterfactuals. IMDB CF (K. et al.) and Cont.Sets refer to the human-authored counterfactuals (Kaushik et al., 2019) and contrast sets (Gardner et al., 2020), respectively. $|D_{\text{train}}|$ shows the total number of training instances, including 8,173 original IMDB-S training examples. Results report mean over 30 different random seeds, with s.d. as a subscript. All models are based on the RoBERTa-base architecture. Best results using auto-generated CFs for training are in boldface. Results for NeuroCFs-1g and NeuroCFs-np are underlined when a one-tailed t-test indicates that their improvements over both Yang et al., 2021 and Wang and Culotta, 2021 are statistically significant ($p < 0.05$).  indicates manually created counterfactuals.


$ D_{\text{train}} $	$ D_{\text{train}}^{\text{CF}} $	Source of CFs	IMDB			SST-2		Out-of-domain		
			Test	CF (K. et al.)	Cont.Sets	Test	PolyJuice CFs	Twitter	Yelp	Amazon
8,173	0	None	93.22 _{0.42}	92.07 _{1.04}	86.85 _{1.06}	90.30 _{0.97}	84.74 _{0.46}	77.94 _{1.72}	94.71 _{0.67}	90.35 _{1.03}
		Yang et al., 2021	91.68 _{0.91}	91.91 _{1.65}	86.69 _{1.76}	89.73 _{1.05}	87.24 _{0.51}	77.03 _{2.20}	93.22 _{1.31}	90.02 _{1.20}
	Wang and Culotta, 2021	92.66 _{0.52}	94.17 _{1.21}	89.41 _{1.50}	89.15 _{1.30}	85.87 _{0.53}	77.62 _{1.67}	94.23 _{0.70}	90.98 _{0.84}	
	NeuroCFs-1g	92.58 _{0.71}	93.35 _{0.86}	88.20 _{1.12}	92.13 _{0.60}	86.63 _{0.55}	78.88 _{1.37}	94.93 _{0.52}	91.80 _{0.72}	
	NeuroCFs-np	92.87 _{0.46}	94.75 _{0.64}	89.99 _{0.94}	<u>92.04</u> _{0.70}	87.64 _{0.57}	<u>78.72</u> _{1.51}	<u>94.76</u> _{0.55}	91.87 _{0.73}	
	 Kaushik et al., 2019	93.09 _{0.46}	96.06 _{0.37}	92.81 _{0.79}	90.99 _{0.82}	88.48 _{0.43}	80.30 _{1.60}	94.52 _{0.81}	91.87 _{0.89}	

Table 3: Results controlling for training data quantity ($|D_{\text{train}}|$), comparing different counterfactual data augmentation approaches. The first row shows a baseline trained without CFs. All other settings are identical to Table 2.

in-domain test performance on IMDB-Test, relative to training without CFs.

Each source of CFs evaluated in Table 2 produces different amounts of training data, D_{train} . To control for training data quantity, we present results with downsampling the training data for uniformity across settings, in Table 3. Surprisingly, even lower amounts of NeuroCFs achieve the best performance compared to other methods of autogenerating CFs. Notably, NeuroCFs-np achieves statistically significant improvements over both Yang et al. (2021) and Wang and Culotta (2021) on every evaluated dataset. These results demonstrate that the performance improvements achieved on OOD sets can be attributed to the quality of the NeuroCFs. App. C.1 provides further results on sentence-level tests.

Table 4 compares our NeuroCFs and CFs from other sources, to the original, across three similarity metrics: BLEU (n -gram = 2) (Papineni et al., 2002), Levenshtein edit distance (Levenshtein et al., 1966), and MoverScore (Zhao et al., 2019). Additionally, Table 4 provides the mean perplexity of generated counterfactuals as measured by GPT-J (Wang and Komatsuzaki, 2021) as well as the Distinct-2 diversity measure (Li et al., 2015). Neu-

roCFs are loose counterfactuals by design, and are therefore farther away from the original sentence; NeuroCFs-np are tighter CFs compared to NeuroCFs-1g. However, NeuroCFs have greater fluency (as evidenced by lower mean perplexity) and offer performance benefits over more similar CFs via minimal edit approaches (Table 2). Moreover, more dissimilar variants, generated without constraints for generation (§4.3), or with alternative concepts (§4.4) also hurt performance.

4 Analysing NeuroCFs

We present further analysis of NeuroCF properties, such as NeuroCFs size (§4.1), and similarity to the original (§4.2), and also ablations of our method (§4.3, §4.4).

4.1 Impact of NeuroCF quantity

In contrast to minimal edit approaches, our approach has the added advantage of producing more than a single NeuroCF for each original example, via NeuroLogic hyperparameter variation. We seek to investigate how the quantity of NeuroCFs for training data augmentation impacts OOD generalization. To investigate the effect of size beyond results in Table 2, we generate more NeuroCFs-np



Source of CFs	BLEU	Levenshtein	MoverScore	Ppl	Distinct-2
 Kaushik et al., 2019	0.74	20	0.70	19.3	0.49
Yang et al., 2021	0.80	8	0.81	29.1	0.56
Wang and Culotta, 2021	0.56	13	0.65	65.6	0.58
NeuroCFs-np	0.50	48	0.46	12.7	0.45
w/ concept-altered	0.43	70	0.38	18.5	0.51
NeuroCFs-1g	0.10	89	0.20	14.1	0.38
w/o constraints	0.03	97	0.07	4.6	0.32

Table 4: Comparing fluency, diversity, and similarity of generated and human () CFs to the original, across various metrics. NeuroCFs are loose counterfactuals by design, and are therefore farther away from the original sentence.

by varying the length penalty in NeuroLogic from 0.1 to 0.7 in increments of 0.2. Among these candidate counterfactuals for each original instance, we augment the training data with the generation with the lowest MoverScore to our initial NeuroCFs-np. This increases the quantity of NeuroCFs-np from 4,732 to 7,489.

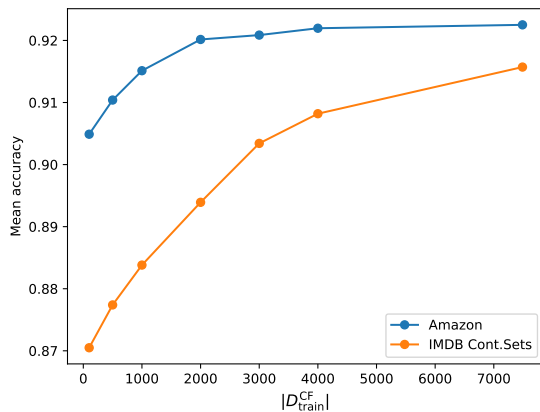


Figure 2: Increasing NeuroCF quantity for training data augmentation improves in-domain performance, while OOD generalization plateaus.

Results in Figure 2 show monotonic increase in accuracy on IMDB contrast sets (Gardner et al., 2020) with NeuroCFs-np size. However, performance on the Amazon OOD set plateaus, suggesting overfitting to the IMDB domain; this echoes the findings of prior work on the efficacy of counterfactuals (Khashabi et al., 2020; Huang et al., 2020; Joshi and He, 2022).

4.2 Impact of NeuroCF Similarity

We investigate the impact of the similarity of NeuroCFs to the original example on sentiment classification performance after augmentation. From the NeuroCFs candidate set described in §4.1, we create two sets of alternative NeuroCFs for each instance: one with the lowest MoverScore (most

dissimilar) w.r.t. the original (NeuroCFs_{loose}) and the other with the highest MoverScore (most similar; NeuroCFs_{tight}).

Table 5 compares these two alternatives via classifier performance across our in-domain and out-of-domain tests. In general, we observe that tighter (i.e., more similar to the original sentence) counterfactuals improve generalization more when evaluated on counterfactual and contrast sets. They also improve out-of-domain generalization, with the exception of the Yelp dataset where both variants result in similar performance. Tighter counterfactuals are more likely to break spurious correlations that help classifiers perform better on in-domain test sets, which may explain why NeuroCFs_{loose} performs better on IMDB Test and SST Test. While NeuroCFs are designed to be loose CFs, these results suggest that higher similarity between the original and its NeuroCF is still important for generalization.

4.3 Impact of Constrained Decoding

Our approach uses a sentiment-steered LM to control the sentiment of the NeuroCFs, and constraint-based decoding to encourage its similarity to the original example. To investigate the impact of constraint decoding, we run an ablation without the use of NeuroLogic, i.e., only using the sentiment steer. Specifically, we use the first token of each original sentence as a prompt and decode from our sentiment experts using beam search with the same hyperparameters as NeuroCFs.

Table 4 compares both variants by their similarity to the original, and Figure 3 compares the performance of training data augmentation with both variants. The use of constrained-based decoding results in substantial performance improvements over the no-constraint baseline across all evaluation sets except the in-domain IMDB test set. This

NeuroCF-Variants	MoverScore	IMDB			SST-2		Out-of-domain		
		Test	CF (K. et al.)	Cont.Sets	Test	CF (PolyJuice)	Twitter	Yelp	Amazon
NeuroCFs _{loose}	0.114	92.50 _{0.59}	93.31 _{0.71}	88.35 _{0.71}	92.26 _{0.56}	86.56 _{0.44}	76.95 _{1.62}	95.01 _{0.42}	91.51 _{0.78}
NeuroCFs _{tight}	0.373	92.24 _{0.68}	93.33 _{0.56}	89.29 _{0.71}	92.23 _{0.55}	86.80 _{0.41}	77.73 _{1.22}	94.93 _{0.28}	92.00 _{0.59}

Table 5: Impact of the similarity of a NeuroCF to the original. NeuroCFs_{loose} are more dissimilar to the original, than NeuroCFs_{tight}, as given by the mean MoverScore. Tighter NeuroCFs result in better performance.

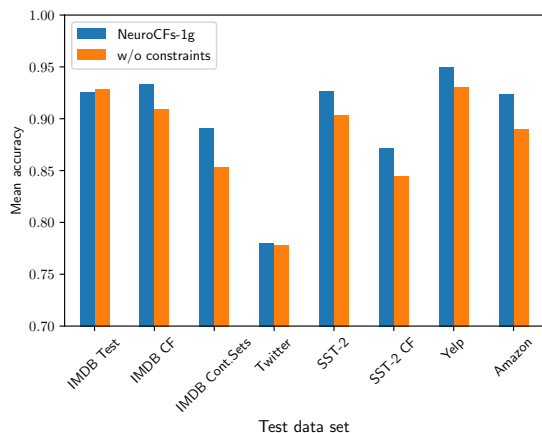


Figure 3: Conceptual constraint-based decoding with NeuroLogic improves performance, as seen by the comparison between training data augmentation with NeuroCFs-np, and their counterparts generated without any constraints. Reported RoBERTa-base accuracy is averaged over 30 random seeds.

highlights the value of using constraints to encourage similarity to the original, thus resulting in a NeuroCF, as opposed to simply a new example of the opposite polarity. These results, along with those from §4.2 indicate the existence of an optimal degree of similarity, which is not as high as minimal edit counterfactuals, and not as low as constraint-free counterexamples.

Initial experiments further point to the value of ConceptNet constraints, as opposed to nominal constraints; the former results in more similar NeuroCFs (see App. §C.2 for details).

4.4 Leveraging ConceptNet for alternative constraint sets

Our use of COCO-EX for identifying concept constraint sets provides a link between each of our constraints and a node in ConceptNet. We explore whether the structured knowledge contained in ConceptNet can provide alternative constraint sets for NeuroCFs.

For each concept in our original constraint sets, we query ConceptNet for its most similar¹¹ English-

¹¹Via similarity scores calculated over pre-computed Con-

language node in the graph and use the label of this nearest neighbor to replace our original concept constraint (see Appendix C.3 for examples). Table 6 compares the performance of a RoBERTa-base classifier trained on NeuroCFs-np, and their counterparts produced by alternative conceptual constraints derived from ConceptNet, and a combined set of NeuroCFs produced by both the original and concept-altered constraints. We observe that further increasing the size of our CFs using concept-altered NeuroCFs increases performance on in-domain CF test sets while retaining performance on OOD test sets. While this pilot shows promising results, we leave a systematic investigation into ConceptNet knowledge to create counterexamples for data augmentation, to future work.

4.5 Can NeuroCFs be used for evaluation?

Inspired by the success of NeuroCFs for training data augmentation, we further investigate if these can be used as a challenge set for evaluation (Rudinger et al., 2018). However, before deploying them as test sets, we need to first verify that NeuroCFs indeed alter the target label, as intended by the sentiment steering process (§2.2). We randomly select 50 NeuroCFs, as well as CFs from baseline approaches, to evaluate whether they successfully steered the sentiment of the original example.¹² Results show that NeuroCFs-np and NeuroCFs-1g are more successful in steering sentiment compared to the baseline approaches; however, only about 50% of the resulting NeuroCFs-np actually result in sentiment change; see further discussion in App. C.4. Hence, we cannot reliably use generated counterfactuals for evaluation. Future work might investigate manually labeling NeuroCFs for use as challenge sets, following Wu et al. (2021).

5 Related Work

Counterfactual data augmentation is emerging as a viable solution for improving model robustness

ceptNet Numberbatch embeddings.

¹²To ensure fairness, the source of the counterfactual as well as the intended label was kept hidden during validation.

NeuroCFs Constraints	IMDB			SST-2		Out-of-domain		
	Test	CF (K. et al.)	Cont.Sets	Test	CF (PolyJuice)	Twitter	Yelp	Amazon
Original	92.66 _{0.46}	95.03 _{0.47}	90.85 _{0.84}	92.27 _{0.39}	88.35 _{0.41}	78.80 _{1.22}	94.51 _{0.87}	92.24 _{0.71}
Concept-altered	92.85 _{0.63}	95.26 _{0.78}	90.55 _{1.10}	91.94 _{0.49}	87.98 _{0.65}	78.53 _{1.51}	94.32 _{0.80}	92.01 _{0.71}
Original + concept-altered	91.83 _{0.65}	96.04 _{0.46}	91.86 _{1.06}	91.38 _{0.42}	88.35 _{0.43}	78.57 _{1.63}	93.90 _{0.92}	92.03 _{0.63}

Table 6: Impact of concept-altered constraint sets created from ConceptNet on classifier performance

towards spurious correlations (Geirhos et al., 2020). In previous sections, we present comparisons to various minimal edit approaches for producing counterfactuals (Kaushik et al., 2019; Wang and Culotta, 2021; Yang et al., 2021; Wu et al., 2021; Gardner et al., 2020), either manually or automatically. Our approach steers away from minimal edits, as well as manual intervention for creating counterfactuals.

Beyond sentiment classification, this approach has been employed for tasks such as question answering (Paranjape et al., 2022), fairness in social computing (Sen et al., 2021), and natural language inference (Glockner et al., 2018). Most work focus on minimal edits of training instances via small perturbations to the causal features, via manually editing instances. Madaan et al. (2021) introduce a controlled text generation approach to create counterfactuals containing specific attributes, but focus on applications to debiasing and evaluation rather than our objective of training data augmentation. Hu and Li (2021) propose a structural causal model for combing attribute-conditional generation and text attribute transfer (i.e., minimal edits), but similarly produce counterfactuals for different purposes than ours. Ross et al. (2022) automate contrast sets (Gardner et al., 2020) for question answering, dependency parsing and relation extraction, via training a generator with semantic control codes; however, their method requires the user to specify what changes in the original sentence are desired.

Beyond Counterfactuals: Srivastava et al. (2020) collect human annotations for common-sense reasoning behind examples, in a robust optimization setting to minimize worst-case loss, without explicitly collecting counterfactuals. Ribeiro et al. (2018) demonstrate how state-of-the-art models are vulnerable to semantically-equivalent adversarial examples constructed from a rule-based method. Ribeiro et al. (2020) propose Checklists, which contain heuristic edits of the evaluation data instances. Other approaches employ perturbations without creating actual data instances (Veitch et al., 2021).

6 Discussion

We presented an approach to generate NeuroCFs, via sentiment steering and concept-constrained decoding. Training data augmentation with NeuroCFs results in improvement on sentiment classification performance over existing minimal-edit methods, both in and out of domain; even matching human counterfactuals in some cases. We presented several analyses for NeuroCFs, and ablations showing the effectiveness of our approach. While NeuroCFs are loose by design, our analyses indicate the existence of an optimal degree of similarity, which is not as high as minimal edit counterfactuals, and not as low as constraint-free counterexamples.

While this work focused on NeuroCFs for movie reviews only, our results show that training on them transfers to other domains such as product reviews and social media posts for the same sentiment analysis task. Future directions of research might investigate generating NeuroCFs for evaluation, and tasks beyond sentiment classification. Our approach is broadly compatible with tasks for which a language model steer can be trained; future applications of this work could therefore include other NLP tasks where global attributes are available, such as toxicity removal or style transfer. Further, we could consider generating a NeuroCFs neighborhood around individual instances, similar to contrast sets (Gardner et al., 2020).

Acknowledgments

We thank Ximing Lu and Chandra Bhagavatula from the Allen Institute for AI for help with the NeuroLogic decoding codebase,¹³ and members of UW NLP, particularly Alisa Liu and Suchin Gururangan for valuable feedback on earlier drafts of this paper. Additionally, we thank Zev Rivlin and Joscha Bach of Intel Labs for their insights throughout the project. We would also like to thank the anonymous reviewers for their constructive input.

¹³https://github.com/GXimingLu/neurologic_decoding

Limitations

Our approach to generate NeuroCFs is designed specifically for binary sentiment classification in English language only. For generating NeuroCFs, we needed the knowledge of the original example’s sentiment polarity; however, it is possible to produce NeuroCFs for both polarities without knowledge of the original label. Applications to other classification settings might involve the need to train multiple language model steers, which can be challenging in the absence of global labels (for e.g. instance-specific labels in multiple-choice question answering). NeuroCFs might need to be filtered for grammaticality and for steering accuracy for their use beyond training data augmentation. Our approach investigated producing loose counterfactuals at the sentence level; efficient extensions of our approach to paragraph-level transformations were not explored in this work. Throughout this work, we use RoBERTa-base and GPT2-Large architectures; however, there are more powerful architectures which could potentially improve our results.

It is possible that language generated through automatic approaches, and labeled automatically might contain their own annotation artifacts (Gururangan et al., 2018), leading to a different set of spurious biases. Potential harms of generated language include harmful social biases (Bender et al., 2021), which were not investigated in this work. Approaches that involve a human validation phase after data collection (Liu et al., 2022), might be explored in future work to mitigate such harms.

Ethical Considerations

We acknowledge that generated language is susceptible to harmful social biases (Bender et al., 2021) and toxicity (Gehman et al., 2020). We caution practitioners against training models *solely* on model generated data. We do not filter our training data or our generations for toxicity, bias, or offensiveness. Hence, we recommend practitioners interested in using our generations and replicating this work to carefully check the generated content before deployment in any real world application.

Our work uses only publicly available datasets. To the best of our knowledge, these do not contain any explicit information about a user’s identity, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, beyond their reviews on movies and products.

References

- Jacob Andreas. 2020. [Good-enough compositional data augmentation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.
- Maria Becker, Katharina Korfhage, and Anette Frank. 2021. [Coco-ex: A tool for linking concepts from texts to conceptnet](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 119–126.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Jacob Eisenstein. 2022. [Informativeness and invariance: Two perspectives on spurious correlations in natural language](#). In *Proc. of NAACL*.
- Amir Feder, Katherine A. Keith, Emaad A. Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2(11):665–673.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that](#)

- require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.
- William Huang, Haokun Liu, and Samuel R. Bowman. 2020. [Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data](#). In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 82–87, Online. Association for Computational Linguistics.
- Nitish Joshi and He He. 2022. [An investigation of the \(in\)effectiveness of counterfactually augmented data](#). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. [Learning the difference that makes a difference with counterfactually-augmented data](#). In *International Conference on Learning Representations*.
- Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C. Lipton. 2021. [Explaining the efficacy of counterfactually augmented data](#). In *Proc. of ICLR*.
- Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. [More bang for your buck: Natural perturbation for robust question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. [Wilds: A benchmark of in-the-wild distribution shifts](#).
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The winograd schema challenge](#). In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR’12*, page 552–561. AAAI Press.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. [Dexperts: Decoding-time controlled text generation with experts and anti-experts](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. [WANLI: worker and ai collaboration for natural language inference dataset creation](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized bert pretraining approach](#).
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Neurologic decoding:\(un\) supervised neural text generation with predicate logic constraints](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Dip-tikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524.

- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). 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 188–197, Hong Kong, China. Association for Computational Linguistics.
- Bo Pang and Lillian Lee. 2005. [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. [Retrieval-guided counterfactual generation for QA](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, Dublin, Ireland. Association for Computational Linguistics.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. [Do ImageNet classifiers generalize to ImageNet?](#) In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. [Semantically equivalent adversarial rules for debugging NLP models](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. [SemEval-2017 task 4: Sentiment analysis in Twitter](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. [Tailor: Generating and perturbing text with semantic controls](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. [How does counterfactually augmented data impact models for social computing constructs?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Herbert A. Simon. 1954. [Spurious correlation: A causal interpretation](#). *Journal of the American Statistical Association*, 49(267):467–479.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Thirty-first AAAI conference on artificial intelligence*.
- Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. [Robustness to spurious correlations via human annotations](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Antonio Torralba and Alexei A Efros. 2011. [Unbiased look at dataset bias](#). In *CVPR 2011*, pages 1521–1528. IEEE.

- Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. [Counterfactual invariance to spurious correlations in text classification](#). *Advances in Neural Information Processing Systems*, 34.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. [Causal mediation analysis for interpreting neural nlp: The case of gender bias](#).
- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Zhao Wang and Aron Culotta. 2021. [Robustness to spurious correlations in text classification via automatically generated counterfactuals](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14024–14031.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. [Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723.
- Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. [Exploring the efficacy of automatically generated counterfactuals for sentiment analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. [Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance](#). 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 563–578.

A Extended Qualitative Analysis

A larger qualitative analysis is provided in Table 14, which further highlights how NeuroCFs result in more complex changes to the original sentence, and are more successful in sentiment steering than minimal-edit counterfactuals. Minimal edits are at times unable to result in meaningful sentiment flips, and result in reduced grammaticality and pragmatics, producing phrases such as “racism was best” (W&C), and “part in the game” (Y et.al.).

A.1 Examples of cases where a counterfactual was not generated

Table 9 provides examples of sentences for which a NeuroCFs-np was not generated. In these cases, no prefix of the original sentence at least 4 tokens in length was predicted to be neutral. This can be attributed to sentiment-bearing words being present at the start of the sentence.

B Data Augmentation Experimental Setup

B.1 Sentence-level IMDB

We augment the training data of a sentence-level version of this dataset (**IMDB-S**), introduced by Wang and Culotta (2021). Here, the original paragraph-length examples were disaggregated, by splitting the original paragraph into sentences and selecting those which contain keywords highly correlated with labels predicted by a binary sentiment classifier. Each sentence inherits its label from the original paragraph, and Wang and Culotta (2021) found that 96.8% of the inherited labels were accurate based on a manual evaluation of 500 samples.

B.2 Dataset Sizes

Dataset	$ D_{\text{train}} $	$ D_{\text{test}} $
IMDB-S	8173	2245
IMDB-S CF	—	2381
IMDB	—	488
IMDB CF (Kaushik et al., 2019)	—	488
IMDB Cont.Sets (Gardner et al., 2020)	—	488
SST-2	—	1821
SST-2 CF (Wu et al., 2021)	—	3014
Twitter	—	4678
Yelp	—	38000
Amazon	—	941534

Table 7: Size of datasets used in experiments

Table 7 provides details on the size of the datasets used in our experiments. All datasets consist of English language text which we used without

Source of CFs	$ D_{\text{train}} $	Mean Training Time
None	8,173	641.90
Yang et al., 2021	10,376	773.79
Wang and Culotta, 2021	10,744	827.94
NeuroCFs-np	12,905	746.86
NeuroCFs-1g	15,437	927.50
† Kaushik et al., 2019	16,679	788.13

Table 8: Average time (in seconds) to train RoBERTa-base on various sets of counterfactuals measured across 30 random seeds

modification. For training our baselines, Wang and Culotta (2021) provided the sentence-level variants for Kaushik et al. (2019)’s counterfactuals, and we apply their method to obtain the sentence-level counterfactuals from Yang et al. (2021).

B.3 Models and Hardware Details

Our sentiment classifier consists of a RoBERTa-base model (Liu et al., 2019) finetuned on various training data setups for a maximum of $10k$ steps using the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 16 and a learning rate of $1e-06$. We evaluate performance every 500 steps on a validation set randomly sampled from 20% of the training data and terminate training early if there is no improvement for 5 consecutive evaluations. All sources of counterfactuals are evaluated using the same hyperparameters and strategy for withholding validation data.

Our experiments were conducted on a Slurm linux cluster with Nvidia RTX 3090 GPUs. We parallelized the generation of NeuroCFs across 32 GPUs in this environment, resulting in a total running time of 75 minutes. Table 8 reports the mean time to train our RoBERTa-base classifier on the various sets of counterfactuals, measured across 30 different random seeds. Each training run for a given source of counterfactuals and seed was conducted on a single GPU. RoBERTa-base has 125M parameters.

C Additional Results

C.1 Evaluating on sentence-level test sets

Table 12 shows the results of all our approaches and baselines on sentence-level test sets.

C.2 Noun Chunk Concepts as Constraints

As detailed in Section 2.1, we form our constraint sets by using COCO-EX to identify meaningful concepts in the original example. To investigate

Candidate prompt	Original sentence
Long , boring ,	Long, boring, blasphemous.
Do something worthwhile ,	Do something worthwhile, anything really.
Awful , despicable ,	Awful, despicable, unpleasant, unhappy, unredeemable saga of a complete Loser.
This is a good	This is a good, dark film that I highly recommend.
I really liked the	I really liked the black and white cinematography.

Table 9: Examples of cases where a NeuroCFs-np was not generated

	Steering Acc.
Yang et al., 2021	0.24
Wang and Culotta, 2021	0.28
NeuroCFs-1g	0.40
NeuroCFs-np	0.46

Table 10: Accuracy of sentiment steering, based on manual evaluation by authors of this work, on 50 randomly sampled IMDB-S train instances for which CFs were available from all approaches. Many generations from each approach were ungrammatical and unpragmatic (see examples in Table 1 and Table 14), and we considered them as incorrectly sentiment-steered.

how our concept-constrained generations differ from those produced by constraint sets derived from nouns, we generated an alternative set of NeuroCFs using constraints consisting of noun chunks identified by spaCy¹⁴. We found that these alternative noun-chunk conceptual NeuroCFs had an average MoverScore of 0.15 w.r.t. their corresponding COCO-EX concept-constrained NeuroCFs, indicating that the use of concepts for constraint formulation produces substantially different counterfactuals than the use of noun chunks for constraints. Moreover, based on the evidence from Table 5, we hypothesize that these alternative concepts might not result in a performance boost.

C.3 Examples of concept-altered constraint sets derived from ConceptNet

Table 13 provides examples of our original NeuroCFs-np and their concept-altered versions after replacing constraints with similar nodes from ConceptNet.

C.4 Evaluating with NeuroCFs

Table 10 shows the steering accuracy of NeuroCFs as well as CFs from baseline approaches, as evaluated by the authors of this work on a sample of 50 randomly selected examples from each. Some

¹⁴<https://spacy.io/>




Source of CFs	$ D_{\text{test}}^{\text{CF}} $	Acc.	$\Delta(\uparrow)$
None	2245	80.46	0.0
 Gardner et al., 2020	4545	67.52	12.68
 Kaushik et al., 2019	2381	77.57	2.63
NeuroCFs-1g	2051	67.63	12.57
NeuroCFs-np	1322	56.81	25.39

Table 11: Classification accuracy of an off-the-shelf sentiment classifier from the Huggingface Transformers library (RoBERTa-base finetuned on the Yelp dataset). Each row indicates an evaluation set comprised of counterfactuals of the original IMDB-S test set (top row), from different sources. $|D_{\text{test}}^{\text{CF}}|$ indicates size of the counterfactual test set.  indicates manually created counterfactuals. Greater the Δ , more challenging the CF test set. However, NeuroCFs-1g and NeuroCFs-np do not possess human-annotated target labels; also see §4.5.


Source of CFs	$ D_{\text{train}} $	IMDB-S	
		Test	CF (K. et al.)
None	8,173	80.46 _{0.55}	75.21 _{0.84}
Yang et al., 2021	10,376	79.68 _{0.50}	77.63 _{0.71}
Wang and Culotta, 2021	10,744	80.25 _{0.42}	77.62 _{0.90}
NeuroCFs-np	12,905	78.31 _{0.53}	80.01 _{0.71}
NeuroCFs-1g	15,437	79.03 _{0.56}	77.87 _{0.77}
 Kaushik et al., 2019	16,679	77.58 _{0.39}	84.27 _{0.46}

Table 12: Evaluation of counterfactual data augmentation on sentence-level test sets; other settings similar to Table 2.

examples of this annotation can be seen in Table 14 in Appendix A and in Table 1.

We report the performance of a RoBERTa-base classifier finetuned on the Yelp dataset¹⁵ using the original IMDB dataset and various CF test sets in Table 11.

¹⁵<https://huggingface.co/VictorSanh/roberta-base-finetuned-yelp-polarity>

Original Sentence	Original Constraints	Constraints w/ Altered Concepts	NeuroCFs-np	Concept-altered NeuroCFs-np
This is one of the worst movies I saw!	(movies)	(citizen kane)	This is one of the funniest movies I have seen in a long time.	This is one of the best movies I've seen in a long time, and it's also a movie that will make you laugh, cry, think and feel a little bit like citizen kane.
It's maybe the worst comedy spoof ever made.	(spoof) ^ (comedy)	(parodied) ^ (comedic)	It's maybe the best spoof comedy I've seen in a long time.	It's maybe the most parodied comedic film I've seen in a long time.
Unlike many modern stories which seem to revel in dark witchcraft, this is simply a magical tale of hocus pocus that is cute, light hearted, and charming.	(hocus pocus) ^ (tale)	(mumbo jumbo) ^ (story)	Unlike many modern stories which seem to revel in dark witchcraft this is simply a tale of hocus pocus and sleight of hand.	Unlike many modern stories which seem to revel in dark witchcraft this is simply a story about mumbo jumbo and a lot of it.
He really just wants to be a good boy, to do the right thing, and to make his brother proud of him.	(brother)	(younger sibling)	He really just wants to be a good boy to do the right thing for his brother, but he just can't do it.	He really just wants to be a good boy to do the right thing, but his younger sibling isn't buying it.

Table 13: Examples from IMDB-S and their corresponding NeuroCFs-np, generated with original and with concept-altered constraints (see §4.4). The prompt (history) used for NeuroLogic decoding is colored orange.

	Label	Review
Original	⊕	A good enough film that unfortunately leaves you a little sad at the end.
W&C.	⊕	A good enough film that luckily leaves you a little sad at the end
Y.et al.	⊕	A good enough film unfortunately leaves you a little sad at the end.
NeuroCFs-1g	??	A film about the end of the world as we know it.
NeuroCFs-np	⊖	A good enough film that unfortunately leaves you a little sad at the end, but it's not a great one.
Original	⊕	Crash tried to show how racism was bad (and Crash actually had a built-in anti Asian bias) and to come at it from a morally superior position.
W&C.	⊖	Crash tried to show how racism was best and crash actually had a built in anti asian bias and to come at it from a morally superior position
Y.et al.	??	Crash tried to show how was bad (and crash actually had a built - in anti asian bias) and to come at it from a morally superior position.
NeuroCFs-1g	??	Crash crash, anti racism position, asian stereotypes, and.
NeuroCFs-np	⊖	Crash tried to show how racism was bad and crash actually had a built-in anti asian bias and to come at it from a morally superior position, it didn't do much of either.
Original	⊕	Many references to the directors obvious love of film history (a great "Bicycle Thief" homage") and some whimsical scenes which work with out being pretentious.
W&C.	⊕	Many references to the directors obvious love of film history a little bicycle thief homage and some whimsical scenes which work with out being pretentious
Y.et al.	⊕	Many references to the directors obvious love of film history (a great " bicycle thief " homage ") and some whimsical which work with out being pretentious.
NeuroCFs-1g	⊖	Many history buffs will love this film, but it's not for directors or actors, and it doesn't have the charm of scenes like this one, in which a bicycle thief references the film.
NeuroCFs-np	⊖	Many references to the directors obvious love of film history and scenes from the classics are made, but the film is so formulaic that it's hard to care about the characters or the bicycle thief.
Original	⊖	The motorcycle is the worst part in the game.
W&C.	⊕	The motorcycle is the good part in the game
Y.et al.	??	The motorcycle is the part in the game.
NeuroCFs-1g	⊕	The best part of the game is the motorcycle chase sequence.
NeuroCFs-np	⊕	The motorcycle is the best part of the film, and it's a game changer.
Original	⊖	Where this weakness is most apparent is in his humor, which is poised and logical, so the joke is dead before it's delivered.
W&C.	⊖	Where this weakness is most apparent is in his humor which is poised and logical so the well is dead before it s delivered
Y.et al.	⊖	Where this is most apparent is in his humor, which is poised and logical, so the joke is dead before it's delivered.
NeuroCFs-1g	⊕	Where the joke is, the humor comes from the characters.
NeuroCFs-np	⊖	Where this weakness is most apparent is in his humor which is poised and logical so the audience doesn't have to guess what the joke is.
Original	⊕	Jackknife is never getting boring as it shows its heroes clear of any typical Hollywood's typical character elements.
W&C.	⊖	Jackknife is never getting interesting as it shows its heroes clear of any typical hollywood s typical character elements
Y.et al.	⊖	Jackknife is getting boring as it shows its heroes clear of any typical hollywood's typical character elements.
NeuroCFs-1g	⊖	Jackknife offers elements of hollywood melodrama, character development and heroes and villains.
NeuroCFs-np	⊖	Jackknife is never getting the character elements hollywood heroes are supposed to have.
Original	⊖	The scenery looks like cheap Theatre.
W&C.	⊕	The scenery looks like <u>expensive</u> theatre
Y.et al.	??	The scenery looks like cheap theatre.
NeuroCFs-1g	⊕	The scenery, the music, and the theatre are all top notch.
NeuroCFs-np	??	The scenery looks like it could have come straight out of a movie theatre.

Table 14: Further qualitative analysis, extending Table 1.