Beyond Model Interpretability: On the Faithfulness and Adversarial Robustness of Contrastive Textual Explanations

Julia El Zini and **Mariette Awad**
Electrical and Computer Engineering Department
American University of Beirut
Beirut, Lebanon
jwe04@mail.aub.edu mariette.awad@aub.edu.lb

Abstract

Contrastive explanation methods go beyond transparency and address the contrastive aspect of explanations. Such explanations are emerging as an attractive option to provide actionable change to scenarios adversely impacted by classifiers' decisions. However, their extension to textual data is under-explored and there is little investigation on their vulnerabilities and limitations.

This work motivates textual counterfactuals by laying the ground for a novel evaluation scheme inspired by the *faithfulness* of explanations. Accordingly, we extend the computation of three metrics, *proximity*, *connectedness* and *stability*, to textual data and we benchmark two successful contrastive methods, POLYJUICE and MiCE, on our suggested metrics. Experiments on sentiment analysis data show that the connectedness of counterfactuals to their original counterparts is not obvious in both models. More interestingly, the generated contrastive texts are more attainable with POLYJUICE which highlights the significance of latent representations in counterfactual search. Finally, we perform the first semantic adversarial attack on textual recourse methods. The results demonstrate the robustness of POLYJUICE and the role that latent input representations play in robustness and reliability.

1 Introduction

With the unprecedented growing deployment of Machine Learning (ML) models in high-stake areas such as law enforcement and medicine, many concerns are raised about their black-box decision-making process. Recently, transparency is becoming a momentous requirement for accountable ML bringing forth the concept of Explainable AI (ExAI). ExAI is witnessing endeavors in all modalities and textual data in particular (Pruthi et al., 2020; Ribeiro et al., 2016; Lundberg and Lee, 2017). However, such explanations might not be sufficient in critical areas where stronger guarantees and more fine-grained explanations are required. Data controllers and subjects pose strong requirements on the *usefulness* aspect of explanations which implies a selective, contrastive and social process (Forrest et al., 2021; Ribera and Lapedriza, 2019). The latter entails an interaction between the explainers and the explainees and human-understandable explanations (Mittelstadt et al., 2019).

To this end, contrastive[1] explanations have seen a surge of interest as a main tool to reach recourse (Ustun et al., 2019; Mothilal et al., 2020; Madaan et al., 2021). Those explanations search for optimally proximate alternative inputs that would result in a different, *usually desired*, prediction. They offer explanations that are tailored to the recipient's beliefs and comprehension capabilities.

Whilst there is a plethora of literature published on recourse methods applied on tabular datasets and computer vision applications (Mothilal et al., 2020; Dosovitskiy and Brox, 2016; Pawelczyk et al., 2021a), little is available on contrastive textual explanations. Specifically, hand-crafted contrastive sets have been employed before to evaluate fairness and robustness of ML models (Garg et al., 2019) by rewriting input instances (Gardner et al., 2020) and defining perturbation functions (Ribeiro et al., 2020) to obtain counterfactual sets. A novel, yet interesting, vein of research is considering an automated counterfactuals generation in NLP (Ross et al., 2021; Wu et al., 2021). The focus of our work is this largely under-explored *targeted* recourse methods for language data and their evaluation schemes. We highlight scenarios where non-recourse methods fall short of the usefulness

Code available at: https://gitlab.com/awadailab/faithful-contrastive-explanations/

[1] Throughout this work, we use the terms *contrastive*, *counterfactual* and *recourse* interchangeably.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1391–1402
December 7-11, 2022 ©2022 Association for Computational Linguistics

aspect of explanation especially due to the blend of syntax and semantics in the words.

Additionally, we consider the assessment schemes, and we target a novel evaluation aspect of the plausibility and attainability aspect of the generated counterfactuals. We argue that counterfactuals should (1) meet textual attainability from a grammatical and semantic perspective, (2) convey connectedness to their original counterparts, and (3) satisfy local algorithmic stability. Accordingly, we extend *proximity*, *connectedness* and *stability*, in the context of *faithfulness*, to textual data and we propose tangible measures to quantify them. We benchmark our metrics on a sentiment analysis task on two famous recourse methods (Wu et al., 2021; Ross et al., 2021). Our results highlight the role that latent representations in (Wu et al., 2021) play in robustness and plausibility. Finally, we present the first study on the resilience of textual recourse methods in the context of adversarial attacks. The study demonstrates a significant improvement in the adversarial robustness of POLYJUICE over MiCE. Our contribution falls under the following categories:

- Surveying textual counterfactuals and highlighting their usefulness over traditional ExAI
- Proposing new evaluation metrics inspired by explanation *faithfulness* and benchmarking contrastive methods, POLYJUICE and MiCE
- Evaluating the robustness of NLP recourse methods through semantic adversarial attacks

Next, we start by motivating the use of textual counterfactuals in Section 2 and highlighting their interconnection to adversarial attacks in Section 2.2. Then, we present the background needed on counterfactuals methods in Section 3 before reporting current evaluation schemes in Section 4. Finally, we extend the *faithfulness* concept to textual data and validate it in Sections 5 and 6 respectively before concluding with final remarks in Section 7.

2 Use cases of Contrastive Explanations

2.1 Favourable Use-cases

Textual ExAI methods interpret outcomes by highlighting segments that support the decision (Ribeiro et al., 2016, 2018) by computing gradients, attention weights, and simpler approximations. Such methods have fundamental impediments.

First, highlighting important input segments fails to specify the contrast between different decision boundaries. Second, the search space of traditional explainability is restricted to the words in the input text. Such methods forsake an integral space of words whose absence from inputs affected the prediction. This assumption thwarts the comprehensiveness and completeness of the explanations. Finally, the fusion of syntax and semantics in a word makes it hard for a practitioner to identify the precise aspect of the word the model is attending to. For instance, the explanation in Figure 1 shows that the sentiment is negative because of the word *slow*. A user is left uninformed on whether the sentence structure (part-of-speech tag, named entity...) or the meaning of the adjective “slow” is driving the model’s decision.

While the first two limitations are shared by general explainability methods; the last limitation is specific to NLP. Recourse methods address these limitations and present additional assets to the model’s transparency. Their underlying design matches the human perception of explanations. In fact, humans inherently understand explanations contrastively (Kumar et al., 2020). A fact-foil contrast can thus introduce a user-centered explanation aspect that complies with the human-in-the-loop drift in AI.

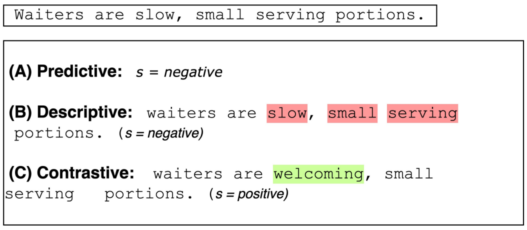
To further highlight the importance of counterfactuals in NLP, we show how they can be leveraged in digital strategy (Boulton and Writer, 2019). We assume an ML model $\mathcal{M}$ that classifies customer comments based on sentiment. We consider three scenarios thereafter illustrated in Figure 1.

(A) Predictive: $\mathcal{M}$ predicts a sentiment $s \in S = \{$ positive, negative, neutral $\}$.

(B) Descriptive: $\mathcal{M}$ predicts a sentiment $s \in S$, with a set of input segments supporting the decision, i.e. through non-recourse explainable AI.

(C) Prescriptive: $\mathcal{M}$ predicts a sentiment $s \in S$, with one (or more) counterfactual inputs that are close to the original input but can change the decision s.

(A) helps the institution in the assessment of their customer satisfaction. A company that hopes to better understand its customers has to resort to **(B)** or **(C)**. Both models analyze user feedback as detractors to identify areas that need improvement. **(B)** highlights the service and the portion size. However, these explanations are not the strong guarantee that strategic planning requires. On the



Waiters are slow, small serving portions.

(A) Predictive: *s = negative*

(B) Descriptive: waiters are slow, small serving portions. (*s = negative*)

(C) Contrastive: waiters are welcoming, small serving portions. (*s = positive*)

Figure 1: Non-recourse Vs. recourse methods

other hand, **(C)** does not stop at highlighting input features that support $\mathcal{M}$'s decision; it generates reviews in a parallel *counterfactual* word that can change the feedback of the user from *negative* to positive. **(C)** can thus prescribe a workable strategy that is more likely to improve the users' feedback.

In addition to that, contrastive methods convey a way to evaluate and improve models and expose bias (Ribera and Lapedriza, 2019; Wu et al., 2021; Sharma et al., 2020; Cheng et al., 2020; Garg et al., 2019; Huang et al., 2020). Contrast sets can also serve as ways to identify general model's vulnerability (Gardner et al., 2020; Artelt et al., 2021) and improve the robustness through contrastive data augmentation (Wu et al., 2021; Qi and Luo, 2020). Contrastive learning is also shown to promote better separation in clustering applications (Zhang et al., 2021) and easier model debugging for non-experts (Ribera and Lapedriza, 2019). (Iter et al., 2020) and (Kiyomaru and Kurohashi, 2021) further prove that pre-training language models with contrastive text improves the discourse coherence between clauses in text.

2.2 *Malignant* Counterfactuals

Can counterfactuals be used as adversarial attacks? To illustrate this, we consider hate speech detection where counterfactuals can be used to make a post get through the "hate speech" check. These modification are *minimal* by design. This brings us to jointly study contrastive explanations (CEs) through the lens of adversarial attacks (AEs) and robustness.

While both methods solve a similar optimization problem, the philosophy behind the optimization of CE and AEs can be conflicting. The former methods compute alternatives that result in a different *desired* prediction. Adversarial alternatives are generally semantically indistinguishable from the original input (Zhang et al., 2019a) whereas counterfactuals are perturbations to actionable features. Interested readers are referred to the work of (Pawelczyk et al., 2021b) that establishes the theoretical and empirical connections between the literature on counterfactual explanations and adversarial examples.

3 Contrastive Textual Explanations

3.1 General Contrastive Theory

Assuming a predictor, potentially non-linear, $f : \mathcal{X} \mapsto \mathcal{Y}$*, an instance* $\boldsymbol{x}^{\boldsymbol{i}} \in \mathcal{X}$ *such that* $f(\boldsymbol{x}^{\boldsymbol{i}}) = y_{fact}$ *and a foil class* y_{foil}*, a counterfactual* $\boldsymbol{x}^i_{cf} \in \mathcal{X}$ *can be computed as:*

$$\underset{\boldsymbol{x}_{cf}}{\arg\min} \quad d(\boldsymbol{x}^i_{cf}, \boldsymbol{x}^{\boldsymbol{i}}) \tag{1}$$

$$\text{subject to} \quad f(\boldsymbol{x}^i_{fl}) = y_{\text{foil}} \tag{2}$$

where $d(.)$ is a distance metric.

This optimization can be also perceived as $\underset{\boldsymbol{x}_{cf}}{\arg\min} \quad L(f(\boldsymbol{x}^i_{fl}), y_{\text{foil}}) + \lambda d(\boldsymbol{x}^i_{cf}, \boldsymbol{x}^{\boldsymbol{i}})$ in the Lagrangian notation, with $l(.)$ denoting a loss function and $\lambda > 0$ is a regularization factor that balances the minimal edit distance and the success in altering the model's decision. Depending on the method, restrictions might complement the above definition.

One line of research deploys gradient-based techniques to achieve a feasible solution (Dhurandhar et al., 2018; Schut et al., 2021). Another track focuses on graph search techniques (Poyiadzi et al., 2020), growing hyper-spheres (Laugel et al., 2017) and integer programming tools (Ustun et al., 2019).

When considering the counterfactual cost or distance $d(.)$, the literature has formed a consensus on the use of the (normalized) l_0 or l_1 norm or any convex combination thereof (Mothilal et al., 2020; Karimi et al., 2020a; Ustun et al., 2019; Wachter et al., 2017). ExAI is witnessing momentum in the adoption of manifold-like distance measures based on adversarial learning (Zhou et al., 2021; Dosovitskiy and Brox, 2016) and Variational Auto-Encoders (VAEs) (Samanta et al., 2020; Joshi et al., 2019). Interested readers are referred to (Stepin et al., 2021) and (Karimi et al., 2020b) that survey state-of-the-art recourse methods.

3.2 Contrastive Methods in NLP

Up until 2021, natural language data was a modality that did not receive enough attention when it comes to recourse methods. Although contrastive sets have been extensively applied to evaluate robustness and fairness (Garg et al., 2019; Huang et al., 2020; Gardner et al., 2020) in NLP, these sets are carefully manufactured as adversarial attacks.

(Jacovi et al., 2021) present contrastive explanations as textual highlights that support one decision against its contrasts. Their approach encodes an input text into a latent vector and applies a linear interpolation where the contrastive direction is found in the latent space. This approach can be viewed as a contrastive search based on the positive pertinent which is not very aligned with the "actionable change" of recourse methods. (Dhurandhar et al., 2018) formulate the pertinent negatives aspect of counterfactuals that are represented by the *missing* features but did not apply their pertinent negatives to natural language data.

Generate Your Counterfactuals (GYC) in (Madaan et al., 2021) is one of the first explicit attempts at automating the generation of textual contrastive explanations. GYC trains a language model to reconstruct the input from the generated counterfactuals and then learns perturbations on the latent space while forcing a proximity constraint.

(Ross et al., 2021) refer to the GYC's proximity criterion as *minimal edits* in their **Mi**nimal **C**ontastive **E**dits framework. To this end, they train a *contextualized* EDITOR to associate edits with task-specific labels by applying masks on input segments that are important for a particular label. EDITOR then serves as a generator by predicting the labels of some masked inputs monitored by binary and beam search to find the optimal maskings.

POLYJUICE (Wu et al., 2021) allows for more edits through negation, replacement, insertion, and deletion for targeted counterfactuals monitored by control codes. Similar to GYC, POLYJUICE relies on a language model to achieve a *fluent* conditional text generation. A filtering layer is added to the process to refine the generated counterfactuals by ignoring the ones that achieve low fluency scores.

The Contrastive Attributed explanations for Text (CAT) of (Chemmengath et al., 2021) inject an attribute prediction layer in the contrastive search process. This layer indicates attributes that the contrast adds to or removes from the given example. Very recently, Malandri et al. (Malandri et al., 2022) develop ContrXT, a **T**ime **Contr**astive model-agnostic e**x**planation framework in lifelong learning settings. ContrXT is not restricted to locally explaining predictions, it rather focuses on the learning process and on how the decision paths of classifiers evolve after retraining. The discussed methods are summarized in Table 1.

4 Evaluation Methods

4.1 Quantitative Evaluation

The most intuitive desiderata for any contrastive explanation are their proximity and ability to change the model's prediction. Both conditions are axiomatically inferred from the problem formulation in Equation 1. In NLP, these conditions are referred to as *minimal distance* and label-flip score respectively. Proximity between $\boldsymbol{x}_{cf}$ and $\boldsymbol{x}$ is measured by word-level Levenshtein distance (Levenshtein et al., 1966) reflecting the edit distance in terms of replacement, insertions and deletions. We draw the reader's attention to the fact that embedding distance measures how similar two vectors are in terms of syntax and semantics (Vylomova et al., 2016) whereas Levenshtein distance reflects the edit distance, or the path to reach counterfactuals. The latter is aligned with the fundamentals of contrastive textual explanations whereas the former is used to measure content preservation. Another way to measure the edit distance is through syntactic trees (Zhang and Shasha, 1989; Wu et al., 2021).

An additional requirement for counterfactuals is the diversity of the generated explanations. Inspired by the Self-BLEU metric of (Zhu et al., 2018), diversity can be measured through the Self-BLEU or Self-BERT (Zhang et al., 2019b) metric between the generated counterfactual samples.

Other requirements that are tailored to natural language are (1) fluency through grammatical correctness and semantic meaningfulness, and (2) content preservation. Fluency can be evaluated by comparing the loss of a particular language model on $\boldsymbol{x}_{cf}$ and $\boldsymbol{x}$ using a pre-trained model (Ross et al., 2021; Morris et al., 2020; Wu et al., 2021). Content preservation can be inferred by latent embedding representations as the cosine similarity between the embeddings of $\boldsymbol{x}_{cf}$ and $\boldsymbol{x}$.

4.2 Qualitative Evaluation

In the social aspects of AI, user studies are ubiquitous in evaluating explainable and fair AI models (Luss et al., 2021; Natesan Ramamurthy et al., 2020; Singh et al., 2018). GYC uses a score to estimate the human judgment of grammatical correctness, plausibility, fluency, sentiment change, and content preservation. Similarly, CAT evaluates human judgment of completeness, sufficiency, satisfaction, and understandability mainly. Instead

Method	Pertinent Negatives	Diversity	Latent Representations	Strategy
Highlights	✗	✗	✓	Linear interpolation on latent space
GYC	✓	✓	✓	Perturbations on latent space and auto-encoder generation
MiCE	✓	✓	✗	Masking input segments and searching for optimal mask combinations
POLYJUICE	✓	✓	✓	Conditional generation with refinement
CAT	✓	✗	✗	Attribute injection in the contrastive search process
Contr-XT	✓	✗	✗	Global and time-sensitive explanations through BDD

Table 1: Summary of existing work on contrastive textual explanations

of surveying human judgment, MiCE's counterfactuals are compared to human edits for overlap, minimality, and fluency. Finally, ContrXT employs crowd-sourcing efforts to evaluate its global explanations, their understandability, and usefulness.

5 Faithfulness Metrics

None of the metrics discussed so far explicitly targets explanation faithfulness that has been studied in non-textual frameworks (Laugel et al., 2019; Pawelczyk et al., 2020). In this work, we redefine *faithfulness* (Laugel et al., 2019), in natural language settings, via three main requirements.

Explanation *faithfulness* entails that counterfactuals should be generated from a "possible" world which is proximate to the starting point specified by the user. This is formalized in two quantitative measures: proximity and connectedness (Laugel et al., 2019) inferring an *attainable* generation of counterfactuals based on the input distribution. Moreover, faithfulness to the explainee engenders local stability of the explainer during the counterfactual generation process.

5.1 Proximity

The contrastive explanation is only useful when presented in terms of plausible means of action. A plausible contrastive text is usually proximate (using a distance notion) to a ground-truth text from the same foil class.

Formally, we consider an instance $\boldsymbol{x}$ belonging to the fact class y_{fact} and its counterfactual $\boldsymbol{x}_{cf}$ belonging to the foil class y_{foil}. Proximity of $\boldsymbol{x}_{cf}$ is measured as the ratio between its distance to $\boldsymbol{x}$ and the minimum distance between $\boldsymbol{x}$ and a ground truth input belonging to the foil class, $\mathcal{X}^{\text{foil}}$.

$$P(\boldsymbol{x}_{cf}) = \frac{d(\boldsymbol{x}, \boldsymbol{x}_{cf})}{\min\limits_{\boldsymbol{x}_{gt} \in \mathcal{X}^{\text{foil}}} d(\boldsymbol{x}, \boldsymbol{x}_{gt})} \tag{3}$$

The notion of proximity is not any different in NLP, except for the computation of the distance metric. This will be discussed later in this section.

Furthermore, we propose the Local Reachability Density (LRD) as a quantitative measure of proximity. LRD reflects how far a point ($\boldsymbol{x_{cf}}$) is from the nearest cluster of points ($\boldsymbol{x}_{gt} \in \mathcal{X}^{\text{foil}}$). Mainly,

$$LRD_k(\boldsymbol{x}_{cf}) = \frac{1}{\sum_{\boldsymbol{x}_{gt} \in \mathcal{N}_k(\boldsymbol{x}_{cf}) \cap \mathcal{X}^{\text{foil}}} \frac{RD(\boldsymbol{x}_{cf}, \boldsymbol{x}_{gt})}{||\mathcal{N}_k(\boldsymbol{x}_{cf})||}} \tag{4}$$

with RD is the reachability distance defined as $RD(i,j) = \max\left(k - \text{distance}(i), d(i,j)\right)$, with $k-\text{distance}(i)$ is the distance from i to its closest neighbor. Higher LRD values are desired as they reflect closer clusters of ground truth data.

5.2 Connectedness

Counterfactual plausibility also engenders a continuous connectedness to the original ground-truth observation. Relying on the topological notion of the path, we borrow the definition of connectedness from (Laugel et al., 2019) as follows:

($\epsilon-$connectedness) $x_1 \in \mathcal{X}$ *is* $\epsilon-$*connected to* $x_2 \in \mathcal{X}$ *if* $f(x_1) = f(x_2)$ *and* $\exists$ *an* ϵ*-chain* $(e_i)_{i<N} \in \mathcal{X}^N$ *between* x_1 *and* x_2 *such that,* $e_0 = X_1$, $e_N = x_2$ *and* $\forall i < Nd(e_i, e_{i+1}) < \epsilon$ $\forall n < N, f(e_i) = f(e)$.

The implementation of the above definition is a compelling problem. We highlight its analogy with density-based clustering (Ester et al., 1996; Laugel et al., 2019) and we adopt the use of DBSCAN algorithm to check whether two texts are connected while setting the *min_points* parameter to 2.

5.3 Stability

This criterion is highly related to robustness where stability requires close counterfactuals for close inputs. Formally, a contrastive explanation method is stable if it satisfies

$$\max_{\boldsymbol{x'} \in \mathcal{B}(\boldsymbol{x},\epsilon)} \frac{d(\boldsymbol{x}'_{cf}, \boldsymbol{x_{cf}})}{d(\boldsymbol{x}, \boldsymbol{x'})} < \epsilon_2 \tag{5}$$

with $\mathcal{B}(x, r)$ is a ball of center x and radius r.

Stability can also serve as an evaluation of the model's resistance to adversarial attacks. (Slack et al., 2021) bring to the front the high sensitivity of recourse methods to insignificant input changes. In fact, gradient-based methods (Wachter et al., 2017) can yield counterfactuals $\boldsymbol{x}_{cf}^{(1)}$ and $\boldsymbol{x}_{cf}^{(2)}$ that are very distant for close inputs $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$.

In summary, *proximity* measures the distance between the encodings of the generated counterfactual x' and the closest instance with the same label in the ground truth data. Connectedness requires x' to be accessible from x along a path consisting of neighboring points with the same label. *Stability* requires the close counterfactuals for close inputs.

Distance measure: The aforementioned notions anticipate a distance measure. Choosing a suitable distance measure is of utmost criticality, especially in NLP, where distances have to reflect syntax and semantics. We use latent space embeddings encoded by language models such as GPT-2 (Radford et al., 2019) and we compute their cosine similarity.

5.4 Discussion

Inadvertently, some of the metrics discussed in Section 4 hint at the faithfulness concept. In general, contrastive attainability can be linked to fluency and grammatical correctness (Wu et al., 2021; Chemmengath et al., 2021; Ross et al., 2021). However, none of the existing work on textual contrastive explanations explicitly addresses *faithfulness*. We draw the reader's attention to the fact that *proximity* is not to be mistaken with the minimal edit distance requirement. The former is the smallest distance between the counterfactual and a ground truth instance belonging to the same class and the latter is a minimal distance between the counterfactual and the original instance.

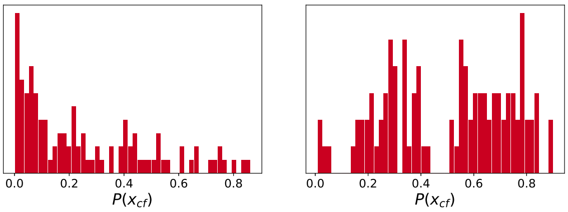
The masking strategy in MiCE can be associated with a search strategy for the topological path. However, this aspect is not explicitly tested. The content preservation concept of (Chemmengath et al., 2021) might be an educated guess at *connectedness* but not a direct measure.

6 Validation of Faithfulness

We consider models with open source code, POLYJUICE, MiCE, and ContrXT mainly. Counterfactuals generated by ContrXT are global which makes *faithfulness* not directly applicable as it evaluates specific (local) explanations. Thus, we consider POLYJUICE and MiCE for our validation. We train both models on the IMDB sentiment analysis task on NVIDIA K80/T4 GPU with 16GB RAM. We consider restaurant reviews for sentiment analysis [2] with 977 validation instances.

6.1 Proximity

We start by evaluating how close the generated counterfactuals are to ground truth data from the foil class. For this purpose, we compute $P(\boldsymbol{x}_{cf})$ of Equation 3 and plot the distribution of its values in Figure 2. One can see a predominance of low proximity scores (< 0.2) in POLYJUICE and an inclination to achieve higher scores with MiCE.



(a) POLYJUICE

(b) MiCE

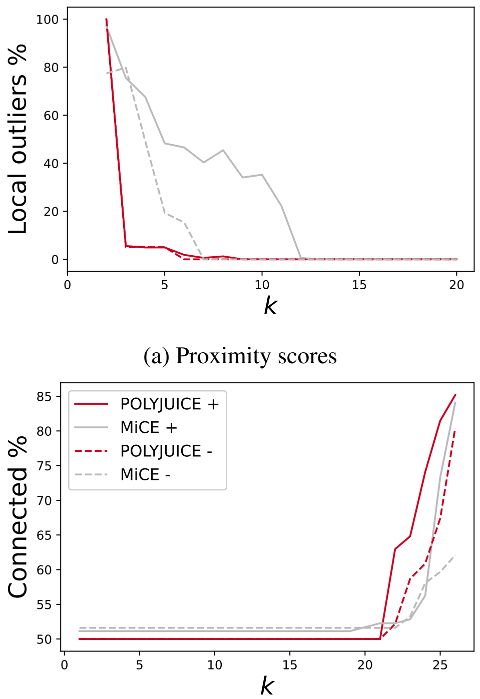
Figure 2: Distribution of $P(\boldsymbol{x}_{cf})$ scores

We further split our validation data according to their foil classes into two categories: positive and negative sentiment foils. For both categories, we compute the outlier factor for the generated counterfactuals, which is inversely proportional to LRD, while changing k and we show the values in Figure 3a. For small k, i.e. strong conditions on outliers, a great deal of the generated counterfactuals, especially with POLYJUICE, are considered outliers. With fair values of k, POLYJUICE drops its generated outliers to nearly zero while some outliers can still be observed with MiCE. Both explanation models are systematic with the foil class being positive or negative sentiments.

6.2 Connectedness

To assess whether the generated counterfactuals are connected to their original factuals, we compute the connectedness score for both explanation

[2] kaggle.com/apekshakom/sentiment-analysis-of-restaurant-reviews



(a) Proximity scores

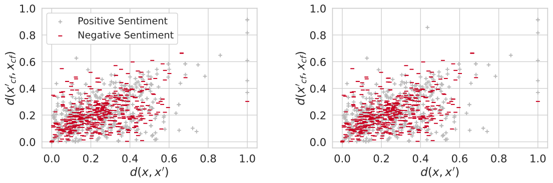
(b) Connectedness scores

Figure 3: Scores while changing the number of neighbors k

models and foil sentiments. The results shown in Figure 3b demonstrate that POLYJUICE and MiCE achieve low connectedness scores when k is small where only half of their generated counterfactuals can be considered connected to the original input. When we loosen the connectedness requirement by increasing k, we notice that more counterfactuals become connected especially with POLYJUICE. For both explanation methods, positive sentiment foil classes seem to achieve higher connectedness scores but the discrepancy between positive and negative sentiments is insignificant with MiCE.

6.3 Stability

We compute $d(\boldsymbol{x}'_{cf}, \boldsymbol{x_{cf}})$ as counterfactual similarity and $d(\boldsymbol{x}, \boldsymbol{x'})$ as input similarity and show how the former measure is scattered in terms of the latter in Figure 4 for POLYJUICE and MiCE. Both plots show that a near-linear correlation governs both models with some high variance. The ratio $\frac{d(\boldsymbol{x}'_{cf}, \boldsymbol{x_{cf}})}{d(\boldsymbol{x}, \boldsymbol{x'})}$ represented by the slope of the linear regression model on the given scatter plots is bounded showing a stability of both explanation algorithms. This can suggest that the non-gradient aspect of the considered contrastive methods yields more robust counterfactuals. The lower variance in POLYJUICE suggests better robustness guarantees. Besides, no significant distinction can be inferred between the two foil categories.



(a) POLYJUICE

(b) MiCE

Figure 4: Scattering of counterfactual similarity with respect to the input similarity. Linear scattering infers local stability.

Finally, we consider more fine-grained stability study, by considering three ranges of input similarities: $d(x, x') < 0.2$, $0.2 \leq d(x, x') < 0.4$ and $0.4 \leq d(x, x') < 0.6$. Figure 5 shows how the counterfactual similarity is distributed for the three considered ranges. Locally, i.e. with input distance < 0.2 POLYJUICE is shown to be more stable on the positive foil class by achieving low distances on the generated counterfactual. MiCE seems to outperform POLYJUICE on the negative foil class. Zooming out, better stability is observed with POLYJUICE for both foil classes.

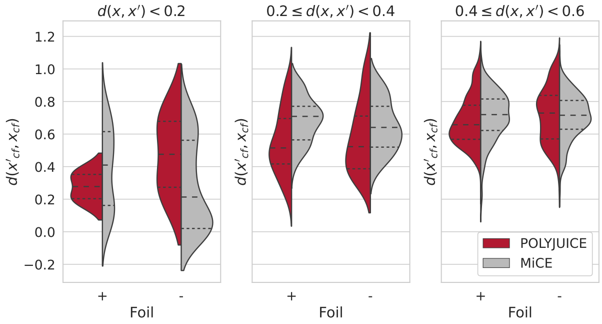
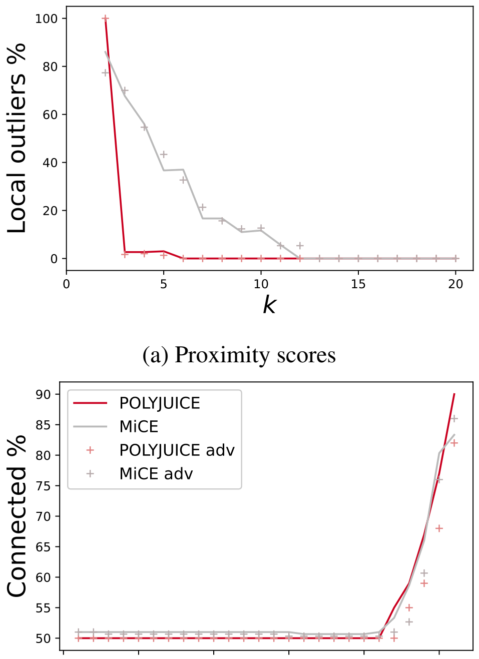


Figure 5: Distribution of the distance between counterfactuals for different input distance ranges

6.4 Adversarial Robustness

We generate adversarial perturbations based on semantic similarity (Morris et al., 2020) on the restaurant reviews. The adversarial inputs are then fed into POLYJUICE and MiCE for a counterfactual generation. Figure 6a demonstrates that the perturbation had no impact on the proximity behavior of POLYJUICE. Markedly, MiCE's counterfactuals became less in-distribution with ground truth data showing questionable robustness to adversarial attacks. The connectedness scores are not affected for both methods as shown in Figure 6b.



(a) Proximity scores

(b) Connectedness scores

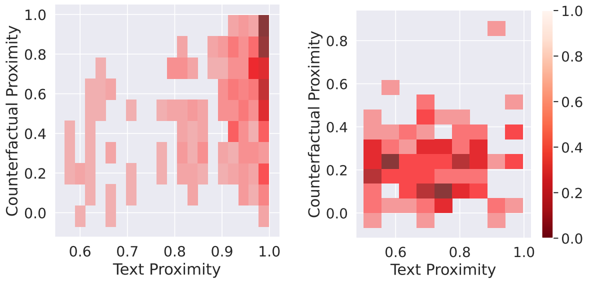
Figure 6: Results with adversarial attacks

Finally, we visualize how the generated counterfactuals are affected when inputs are perturbed. Figure 7 shows the distribution of the cosine similarities between x_{cf} (the counterfactual of the original input, x) and $x_{\text{cf}}^{\text{adv}}$ (the counterfactual of its adversarial counterpart, x^{adv}) with respect to the similarity between x and x^{adv} on a sample of 300 points. POLYJUICE scores higher similarities between counterfactuals showing more robustness to adversarial attacks. Since POLYJUICE does not rely on gradient descent to reach recourse, its results are per the discussion of (Slack et al., 2021) on the problematic behavior of gradient-based counterfactual search on robustness.

While we are aware of the wide range of adversarial textual attacks, we restrict our experiment to semantic similarity and leave the rest for future inspection. We underline that this experiment is different from the attacks discussed in Section 2.2. Rather than attacking the classifier, we perturb its input and feed it into a counterfactual method to study whether the latter is robust to adversarial attacks.

6.5 Comparison to Existing Metrics

We compute existing evaluation metrics, BLEU and Self-BERT mainly, on the generated counterfactuals. On average, POLYJUICE counterfactuals achieve a BLEU score of 0.38 as opposed to a 0.32 score achieved by MiCE. Self-BERT scores were higher, where POLYJUICE and MiCE achieve 0.95 and 0.92 scores respectively.



(a) POLYJUICE

(b) MiCE

Figure 7: Distribution of the cosine similarity of the generated counterfactuals with adversarial attacks

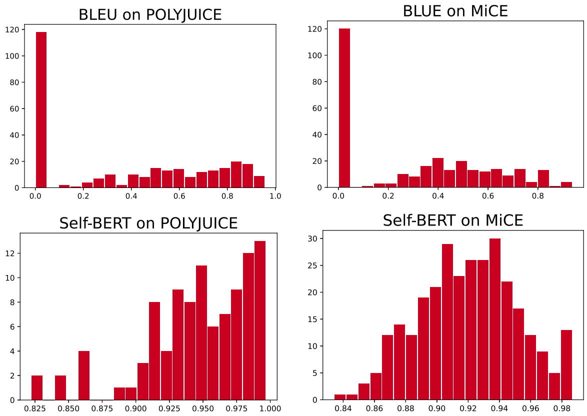


Figure 8: Distribution of the BLEU and Self-BERT scores

The results show a slight improvement of POLYJUICE over MiCE which confirms our findings highlighting again the importance of latent representations. Figure 8 shows the distribution of the scores on the counterfactuals generated by POLYJUICE and MiCE.

6.6 Discussion

The fundamental difference between POLYJUICE and MiCE can be traced to word representations. The former anticipates latent space encodings while the latter operates at the textual level. Hence, we will interpret their faithfulness through the word representation lens.

Proximity results were not consistent. Higher $P(\boldsymbol{x}_{cf})$ scores are reported with MiCE while lower outlier factors are observed with POLYJUICE. One can thus say, that relative to $d(\boldsymbol{x}, \boldsymbol{x}_{cf})$ edits on the textual level achieve higher proximity. Considering a cluster of ground truth inputs with the same class as the counterfactual, POLYJUICE is shown to obey the input distribution in generating contrastive texts. We also call attention to the fluency filtering layer of POLYJUICE which yields better reachability. These results hint at the connection between latent representations and the attainability of generated counterfactuals.

Conversely, connectedness scores do not show any substantial difference. POLYJUICE is shown to be more locally stable and more robust to adversarial attacks. The results make intuitive sense as the distances are computed based on latent representations that are used by POLYJUICE in their contrastive search. Hence, latent representation of words (instead of textual ones) can serve the algorithmic stability of recourse methods.

7 Conclusion

Counterfactual methods go beyond interpretability and offer practical explanations that comply with the social and algorithmic aspects of explainability. In this work we chart the path towards contrastive methods in NLP with a common evaluation scheme inspired by *faithfulness*. We present the limitations of traditional explainability which are leveraged with recourse methods, in NLP mainly.

We further define *faithfulness* of textual explanations and present corresponding computation schemes. Our benchmarks on two famous methods, POLYJUICE and MiCE, show that better algorithmic stability and attainability are achieved in the former, highlighting the importance of latent representation in the counterfactual search strategy. We highlight the vulnerabilities of textual recourse methods against semantic adversarial attacks. Three immediate steps in this line of work are the mitigation of the "unconnected" counterfactuals by posing connectedness constraints on the search strategy, the enhancement of the stability when textual edits are employed, and the investigation of textual attacks on recourse methods.

8 Limitations

While our work addresses one of the limitations of counterfactual textual explanations, *faithfulness* evaluation specifically, it has its own restrictions. First, the connectedness aspect of *faithfulness* is computed based on neighbors sampled from a validation set. This evaluation reflects the plausibility of generated counterfactuals based on the data at hand. A more generalizable plausibility requires sufficiently large validation sets.

One can also argue that *faithfulness* heavily depends on the distance notion which is based on transformers' encodings. Although transformers are state-of-the-art language models that successfully encode syntax and semantics, their performance is crucial for a *faithful* evaluation. Extension to other languages requires careful consideration. This is mainly due to the counterfactual generation process operating differently with different morphologies and so does the distance measure.

Finally, the intriguing relation between counterfactuals and AEs can motivate the use of the former to improve models' robustness against AEs. Informing practitioners of their potential harm is a key responsibility to preventing unfavorable manipulations.

9 Acknowledgment

This work was supported by the University Research Board (URB) and the Maroun Semaan Faculty of Engineering and Architecture (MSFEA) at the American University of Beirut.

References

André Artelt, Fabian Hinder, Valerie Vaquet, Robert Feldhans, and Barbara Hammer. 2021. Contrastive explanations for explaining model adaptations. In *International Work-Conference on Artificial Neural Networks*, pages 101–112. Springer.

Clint Boulton and Senior Writer. 2019. Introducing gail: Great wolf lodge's ai for pinpointing guest sentiment.

Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2021. Let the cat out of the bag: Contrastive attributed explanations for text. *arXiv preprint arXiv:2109.07983*.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2020. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31.

Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29:658–666.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

James Forrest, Somayajulu Sripada, Wei Pang, and George M. Coghill. 2021. Are contrastive explanations useful? pages 9–16. Funding Information: Supported by EPSRC DTP Grant Number EP/N509814/1 ; 2021 SICSA eXplainable Artifical Intelligence Workshop, SICSA XAI 2021 ; Conference date: 01-06-2021.

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.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020a. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020b. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Hirokazu Kiyomaru and Sadao Kurohashi. 2021. Contextualized and generalized sentence representations by contrastive self-supervised learning: A case study on discourse relation analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5578–5584.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. In *ICML Workshop on Human in the Loop Learning (HILL 2019)*.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*.

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.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. 2021. Leveraging latent features for local explanations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1139–1149.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan 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.

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, Navid Nobani, and Andrea Seveso. 2022. Contrxt: Generating contrastive explanations from any text classifier. *Information Fusion*, 81:103–115.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. 2020. Model agnostic multilevel explanations. *Advances in neural information processing systems*, 33:5968–5979.

Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 2021a. Carla: a python library to benchmark algorithmic recourse and counterfactual explanation algorithms. *arXiv preprint arXiv:2108.00783*.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132.

Martin Pawelczyk, Shalmali Joshi, Chirag Agarwal, Sohini Upadhyay, and Himabindu Lakkaraju. 2021b. On the connections between counterfactual explanations and adversarial examples. *arXiv e-prints*, pages arXiv–2106.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793.

Guo-Jun Qi and Jiebo Luo. 2020. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, volume 2327, page 38.

Alexis Ross, Ana Marasović, and Matthew E Peters. 2021. Explaining nlp models via minimal contrastive editing (mice). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852.

Soumitra Samanta, Steve O'Hagan, Neil Swainston, Timothy J Roberts, and Douglas B Kell. 2020. Vaesim: a novel molecular similarity measure based on a variational autoencoder. *Molecules*, 25(15):3446.

Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172.

Chandan Singh, W James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, 34.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.

Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *ACL (1)*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen Mckeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019a. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yingbo Zhou, Pengcheng Zhao, Weiqin Tong, and Yongxin Zhu. 2021. Cdl-gan: Contrastive distance learning generative adversarial network for image generation. *Applied Sciences*, 11(4):1380.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.