AMERICAN UNIVERSITY OF BEIRUT

THEORETICAL GUARANTEES OF CONTRASTIVE LEARNING IN A NOVEL EXPLAINABLE AI METHOD AND A DEEP FAIRNESS EVALUATION FRAMEWORK

by
JULIA WAJDI EL ZINI

A dissertation
submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
to the Department of Electrical and Computer Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
December 2022

AMERICAN UNIVERSITY OF BEIRUT

THEORETICAL GUARANTEES OF CONTRASTIVE LEARNING IN A NOVEL EXPLAINABLE AI METHOD AND A DEEP FAIRNESS EVALUATION FRAMEWORK

by

JULIA WAJDI EL ZINI

Approved by:

Ali Chehab

Dr. Ali Chehab, Professor — Committee Chairperson

Electrical and Computer Engineering

Mariette

Mariette (Jan 26, 2023 12:26 GMT+2)

Dr. Mariette Awad, Associate Professor — Advisor

Electrical and Computer Engineering

Rabih Jabr (Jan 21, 2023 22:20 GMT+2)

Dr. Rabih Jabr, Professor — Member of Committee

Electrical and Computer Engineering

Shady Elbassuoni (Jan 21, 2023 15:54 GMT+2)

Dr. Shadi Elbassuoni, Associate Professor — Member of Committee

Computer Science

Prasenjit Mitra

Dr. Prasenjit Mitra, Professor — Member of Committee

Penn State College of IST

Mykola Pechenizkiy

Mykola Pechenizkiy (Jan 22, 2023 02:34 GMT+1)

Dr. Mykola Pechenizkiy, Professor — Member of Committee

Eindhoven University of Technology

Dr. Carlos Castillo, Professor — Member of Committee

Universitat Pompeu Fabra

Date of dissertation defense: December 19, 2022

AMERICAN UNIVERSITY OF BEIRUT

DISSERTATION RELEASE FORM

Student Name: El Zini Julia Wajdi

Last First Middle

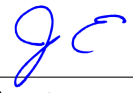
I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my dissertation; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

___ **As of the date of submission of my dissertation**

_x_ **After 1 year from the date of submission of my dissertation .**

___ **After 2 years from the date of submission of my dissertation .**

___ **After 3 years from the date of submission of my dissertation .**



Signature

January 26, 2023

Date

Acknowledgements

My sincere gratitude goes to all my committee members without whom my thesis wouldn't have been shaped the way it is. Prof. Rabih Jabr, Prof. Prasenjit Mitra, Prof. Mykola Pechenizkiy, and Prof. Carlos Castillo, your valuable comments stirred me in the right direction and hugely contributed to the maturity of my thesis.

Prof. Shady Elbassuoni, you have been an integral part of building the scientist I am by teaching me my first machine learning course and making me fall in love with it. I am also extremely grateful to Prof. Ali Chehab for continuously believing in me and supporting me on a personal and professional level.

I would like to extend my sincere thanks to my advisor, Prof. Mariette Awad, who pushed me beyond my limits and made sure I am fulfilling my potential. I am grateful for all the support you provided throughout the pandemic, financial crisis, and Beirut Blast. You never stopped being a great listener and a big support.

Many thanks to my friends and colleagues for the great emotional support, the unforgettable memories, and all the trust and faith in me. Special thanks to Mohamad Mansour for the insightful discussions and his outstanding efforts and hard dedication during the implementation of CEnt.

Last but not least, I am deeply indebted to my parents. My father, Wajdi, thank you for teaching me patience and resilience by fighting Multiple Sclerosis every day. My mother, Ghounwa, thank you for teaching me how to dream and giving me the strength to pursue my goals. My little sister, Najwa, your support for me is just priceless; a ton of thanks for endlessly listening to me and always having my back.

ABSTRACT

OF THE DISSERTATION OF

Julia Wajdi El Zini for Doctor of Philosophy

Major: Electrical and Computer Engineering

Title: Theoretical Guarantees of Contrastive Learning in a Novel Explainable AI Method and a Deep Fairness Evaluation Framework

Given the social implications of autonomous systems in high-stake areas, recent years have witnessed an outpouring of research on designing explainable and fair AI models. In this work, we consider the intersection of contrastive learning with explainable AI and fairness evaluation schemes. Current methods that provide contrastive explainability do not simultaneously satisfy model-agnosticism, immutability, semi-immutability, and attainability constraints. In the fairness framework, existing metrics rely on statistical and causal tools that do not cover all bias cases and do not leverage advances in contrastive learning.

To this end, we present CEnt, a **C**ontrastive **Ent**ropy-based explanation method, to locally contrast the prediction of *any* classifier. CEnt generates contrastive examples and visual contrasts that achieve better proximity rates than existing methods without compromising latency, feasibility, and attainability.

We utilize contrastive sets to devise a novel individual fairness evaluation technique that respects attainability and plausibility by relying on a manifold-like distance metric. Inspired by counterfactual ExAI, we suggest three metrics to evaluate the faithfulness of our metric and we study its interconnection with attainability and plausibility. We demonstrate the effectiveness of our method at detecting bias cases missed by other metrics that do not always satisfy faithfulness requirements.

Furthermore, we extend our fairness metric to textual settings by developing a local method to detect bias cases in textual settings with little reliance on existing ontologies. Our evaluation method computes the statistical mutual information and the geometrical inter-dependency with the sensitive information embedding to evaluate the fairness of a classifier. Likewise, we extend contrastive faithfulness guarantees to natural language by relying on transformers' encodings.

Lastly, we devise a novel mitigation strategy that operates in the latent space by encouraging a classifier to have the same outcome when the latent representation is perturbed with a sensitive direction. Our strategy is effective at diluting, even removing, bias in classifiers without compromising performance.

Our work motivates follow-on research in the fields of contrastive explainable AI, bias detection, and mitigation in deep networks. Generative models can be employed to improve the privacy guarantees of our techniques and enhance the quality and plausibility of the generated contrastive examples.

Table of Contents

ACKNOWLEDGEMENTS **1**

ABSTRACT **2**

ABBREVIATIONS **12**

1 Introduction **14**

2 Background **21**

2.1 Explainable AI: Terminology 21

2.1.1 The *"when"* of Explainability 21

2.1.2 The *"what"* of Explainability 22

2.2 Fairness Terminology 24

3 Related Work **25**

3.1 Objective I: Contrastive Learning for Explainability 26

3.1.1 Explainable AI 26

3.1.2 Counterfactual Explainability 29

3.2 Objective II: Faithful Fairness Evaluation (FCF) 32

3.2.1 Existing Fairness Definitions 32

3.2.2 Distance Measures and Faithfulness 35

3.3 Objective II: Fairness Evaluation in Textual Settings 36

3.3.1 Fairness Evaluation in Textual Classifiers 36

3.3.2 Evaluation of Contrastive Examples in NLP 38
3.4 Objective III: Bias Mitigation 40

4 Objective I: Entropy-based Contrastive Explanations 42
4.1 Problem Statement 43
4.2 Methodology 44
4.2.1 Minimizing Locality-aware Fidelity Loss 44
4.2.2 Minimizing Counterfactual Cost Through Graph Search 46
4.2.3 Complexity 47
4.3 Results 48
4.3.1 Experimental Setup 48
4.3.2 CEnt on Numerical Datasets 50
4.3.3 Derivation of Visual Contrasts 53
4.3.4 Textual Vulnerabilities Detection 56
4.4 Research Directions 58

5 Objective II: Faithful Contrastive Fairness 60
5.1 Problem Statement 61
5.2 Contrastive Fairness 62
5.2.1 Latent distance 63
5.2.2 Computation of FCF 64
5.2.3 Relation to Existing Notions 66
5.2.4 Extension to Group Fairness 67
5.3 Faithfulness Guarantees 68
5.3.1 Stability 68
5.3.2 Proximity 69
5.3.3 Connectedness 69
5.4 Results 70

5.4.1 Synthetic Experiment Setup . . . 70
5.4.2 Synthetic Experiment Results . . . 71
5.4.3 Real World Experiment Setup . . . 73
5.4.4 Real World Experiment Results . . . 74
5.4.5 Comparison to Existing Metrics . . . 75
5.4.6 Faithfulness Results . . . 76
5.4.7 Impact of VAE Architecture . . . 77
5.5 Research Directions . . . 80

6 Objective III: Extension of Contrastive Fairness and Faithfulness Evaluation to Textual Settings 82

6.1 Problem Statement . . . 84
6.2 Methodology . . . 85
6.2.1 Sensitive Attribute Information . . . 85
6.2.2 Sensitive Dependence Evaluation . . . 86
6.3 Results . . . 87
6.3.1 Experimental Setup . . . 88
6.3.2 Computation of the Sensitive Direction . . . 89
6.3.3 Implementation Details . . . 89
6.3.4 Dataset Bias . . . 91
6.3.5 Detected Bias . . . 92
6.3.6 CoFE in the Fairness Evaluation Realm . . . 92
6.4 Faithfulness of Textual Contrasts . . . 94
6.4.1 Proximity . . . 95
6.4.2 Connectedness . . . 96
6.4.3 Stability . . . 96
6.4.4 Adversarial Robustness . . . 97
6.4.5 Comparison to Existing Metrics . . . 99

6.4.6 Discussion . . . 100
6.5 Research Directions . . . 101

7 Objective IV: Latent Bias Mitigation **103**
7.1 Problem Statement . . . 104
7.2 Methodology . . . 104
7.2.1 Regularizationn in General Settings . . . 105
7.2.2 Latent Augmentation in Textual Settings . . . 105
7.3 Results . . . 106
7.3.1 Experimental Setup . . . 106
7.3.2 General Settings . . . 107
7.3.3 Textual Debiasing . . . 108
7.3.4 Fairness-Accuracy Trade-off . . . 110
7.3.5 Comparison to Existing Work . . . 111
7.4 Research Directions . . . 112

8 Conclusion and Future Directions **114**

APPENDIX **117**

Bibliography **122**

Illustrations

3.1 Importance of manifold-like distance metric in neighborhood sampling . . . 36

3.2 LIME explanations of stereotypes on gender prediction. Pronouns revealing gender are hidden. Orange (blue resp.) indicates female (male resp.). . . . 38

4.1 Overview of CEnt . . . 45

4.2 CEnt results averaged on four numerical datasets . . . 51

4.3 Distribution of CEnt's proximity scores across different contrastive methods on the four numerical datasets for LR . . . 52

4.4 Distribution of CEnt's proximity scores across different contrastive methods on the four numerical datasets for ANN . . . 53

4.5 Visual contrast with CEnt on MNIST (red represents PPs and green represents PNs) . . . 55

4.6 MNIST explanations (LIME highlights pixel relevance where red is positive and green is negative, CEM and CEnt highlight contrasts where red is PP and green is PN) . . . 56

4.7 Visual contrast with CEnt on Fashion MNIST (red represents PPs and green represents PNs) . . . 57

4.8 CEnt on an instance of the 20 newsgroup data . . . 58

5.1 Illustration of Algorithm 1. The neighborhood is computed in the latent space $\mathcal{Z}$ and mapped back to the original space $\mathcal{X}$. For illustration purposes, $d = d_2 = 2$. 64
5.2 Synthetic experiment with original data distribution and the generated counterfactuals 71
5.3 Latent space visualization of the neighborhood 72
5.4 Fairness distribution following a log scale for the y-axis 72
5.5 Stability score of FCF on the German and adult datasets 76
5.6 Proximity scores (low numbers of local outliers are desirable) 77
5.7 Connectedness scores (low %not connected scores are desirable) . . . 78
5.8 FCF group fairness scores when different VAEs are used. The green region represents the $[\sigma, \frac{1}{\sigma}]$ range indicating fair treatment. 79
5.9 Impact of VAE architecture on faithfulness 80

6.1 CoFE workflow 86
6.2 CoFE bias scores in the bios dataset 89
6.3 CoFE bias scores in toxicity dataset 91
6.4 CoFE bias cosine scores on the toxicity dataset 93
6.5 CoFE bias on bios data when raw and gender-neutralized biographies are used for training 93
6.6 Distribution of the number of counterfactuals generated by MiCE for each input 94
6.7 Distribution of the $P(\boldsymbol{x}_{cf})$ scores 95
6.8 Scores while changing the number of neighbors k 96
6.9 Scattering of counterfactual similarity with respect to the input similarity. Linear scattering infers local stability. 97
6.10 Distribution of the distance between counterfactuals for different input distance ranges 98

6.11 Proximity and connectedness results with adversarial attacks on textual contrastive examples 98
6.12 Distribution of the cosine similarity of the generated counterfactuals with adversarial attacks 99
6.13 Distribution of the BLEU and Self-BERT scores on the generated counterfactual textual examples 100

7.1 FCF pre- and post- bias mitigation 108
7.2 CoFE bias cosine scores for pre- and post-debiasing where higher cosine scores infer discrimination) 109
7.3 CoFE MI scores for pre- and post-debiasing (Higher MI indicates discrimination) 110
7.4 Example from bios data with LIME explanations highlighted in red (intense means higher correlations) 111
7.5 Accuracy and CoFE with augmentation ratio 111

1 Distribution of FCF following log scale for the y-axis (higher scores indicate unbiased decisions) 121

Tables

3.1 Summary of existing contrastive explanation methods based on their underlying assumptions (gradient-based approaches are not model-agnostic) and on whether they handle immutable, semi-immutable, and categorical features. We also highlight methods that can generate diverse counterfactual examples (CEs) and those whose CEs are generated or selected from the training set. 32

4.1 Accuracies on the original model f and the DT g 51

5.1 Counterfactual age cost for the German credit scoring dataset 73

5.2 Number of individuals discriminated against according to FCF 74

5.3 Group fairness metrics along with FCF on the Adult dataset where underlined scores indicate discrimination with $\sigma = 0.9$ 75

6.1 Pairs of sensitive attributes 89

6.2 CoFE bias vs. equality of opportunity and equalized odds metrics . . 90

7.1 Group fairness metrics along with FCF on the Adult dataset where underlined scores indicate discrimination with $\sigma = 0.9$ 107

7.2 CoFE bias vs. equality of opportunity and equalized odds metrics pre- and post- latent mitigation 113

Abbreviations

ANN	Artificial Neural Network
BoW	Bag-of-words
CE	Contrastive Example
CEnt	Contrastive Entropy based explanations
CF	Counterfactual Fairness
CNN	Convolutional Neural Network
CoFE	Contrastive Fairness Evaluation
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DL	Deep Learning
DP	Demographic Parity
DT	Decision Tree
EqOpp	Equality of Opportunity
ExAI	Explainable AI
FCF	Faithful Contrastive Fairness
FP	False Positive
FTU	Fairness Through Unawareness
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
IF	Individual Fairness
LOF	Local Outlier Factor

LR	Logistic Regression
MI	Mutual Information
MSE	Mean Squared Error
NLP	Natural Language Processing
NN	Neural Network
PCA	Principal Component Analysis
PN	Pertinent Negative
PP	Pertinent Positive
PredP	Predictive Parity
SA	Sensitive Attribute
StatP	Statistical Parity
SVM	Support Vector Machine
TP	True Positive
VAE	Variational Auto-Encoder

Chapter 1

Introduction

Ever since their introduction, Deep Learning (DL) models are revolutionizing several fields ranging from computer vision [1], [2] and machine translation [3]–[5] to question answering [6]–[8], generative models [9]–[11] and complex decision making tasks [12], [13]. AI-powered systems that mainly use DL models can make very accurate predictions or decisions on a wide range of tasks. However, *how credible and trustworthy these predictions are if the reasoning behind them is a highly non-linear enigma that cannot be easily deciphered?*

The black-box nature of DL models gave rise to several criticisms of their non-transparent predictions that are not easily traceable by humans. Relying more and more on AI for decisions requires ethical decisions that are free from unjust biases. The need for responsible AI systems that are transparent, explainable, and accountable is more pronounced than the need for accurate smart systems in high-stakes applications. Moreover, transparency is not only needed on the prediction level; some situations require DL models that achieve transparency on the learning level. For instance, DL models which learn from curated datasets might engender bias [14], [15] that is not easy to detect due to the black-box nature of DL. Hence, interpreting such models, on the encoding level, is crucial to ensure fair treatment of individuals.

These concerns led The US Federal Trade Commission to issue new guidelines requiring AI systems to be open, explainable, and fair. Moreover, the General Data Protection Regulation (GDPR) of the European Union mandates transparency for algorithms and fair representation and treatment in AI systems. Whether or not they operate in the European Union, industries that develop and use data-driven systems are moving into ensuring these regulations. That being the case, data and algorithmic accountability witnessed explosive growth mainly nurtured by the invasive use of autonomous systems and the regulations imposed by legal institutions on data and smart processes.

Accordingly, researchers are extensively engaged in the fields of accountability, fairness, and explainability [16], [17]. This is reflected in developing methods to explain AI decisions and learned representations for different data types. Additionally, researchers are working on providing fairness definitions and bias detection methods in numerous applications. This is mostly accompanied by several techniques to neutralize learned representations and mitigate bias in decision-making systems.

Looking at existing approaches in Explainable AI (ExAI), several methods are proposed to provide post-hoc interpretations by looking for input features, or sets of features, that influence the model's decision [18]–[20]. However, practitioners and data subjects pose strong requirements on the *usefulness* aspect of explanations which implies a selective, contrastive, and social process [21], [22]. This aspect entails an actionable plan which serves as a *constructive feedback* when the prediction is not favorable. As an example, a loan applicant is more interested in how to get their applications accepted rather than the *reason* behind the rejection [23]. To this end, ExAI is witnessing a new vein of methods explaining decisions through contrastive learning motivated by the seminal work of [23]. These recourse methods (a.k.a counterfactual) search for a proximal input that can alter the prediction. They often employ causal graphs, gradient-descent, discriminative, and evolutionary

algorithms to generate Contrastive Examples (CEs) while satisfying feasibility constraints [24]–[26]. Apart from being able to contrast the outputs of various models, the most wanted desiderata of counterfactuals are plausibility, attainability, and diversity. While contrasting the output is successfully achieved by all methods, other requirements partake in a trade-off and are rarely simultaneously satisfied. For instance, some methods violate constraints [23], others do not always yield attainable counterfactuals [27] or output a unique CE based on a proximity measure [28].

Similarly, focusing on fairness in AI literature, one could compile more than twenty different definitions and metrics of fairness that were proposed over the course of the past few years [29]. The simplest and most intuitive definitions are based on the predicted and actual outcome for the different demographic distributions of subjects such as group fairness, statistical parity, and equalized odds [29]. Other definitions are not statistical, they rely instead on causal reasoning [30], [31]. These definitions do not only differ from a theoretical perspective; their outcome might also significantly disagree [32], [33]. In addition, the presented definitions cannot all be satisfied at the same time as proved in [34]. Despite their potential, there are discrimination cases that existing fairness metrics cannot directly detect. Those cases are (1) non-tabular settings where protected (sensitive) attributes are not explicitly reported, (2) situations where sensitive attributes are a *legacy*, (3) the comparison of the individual to an un-attainable counterpart, and (4) the reliance on distance metrics that do not reflect real-world similarity-based measures. On the other hand, de-biasing classifiers remains a challenge especially when it currently relies on the availability and the quality of analog associations and pre-defined parallels that are specific to sensitive attributes [35], [36].

In this work, we extend contrastive learning to resolve the shortcomings of existing ExAI methods, fairness evaluation, and bias techniques. Our contribution is four-fold. First, we design a **C**ontrastive **Ent**ropy-based Explainability method,

CEnt, under feasibility, immutability and semi-immutability constraints while satisfying proximity and user-defined costs. Given an observation x, CEnt samples k local neighbors of x based on manifold-like distance approximated by Variational Auto-Encoders (VAEs). Then, CEnt approximates a black-box machine learning model by a decision tree in the local neighborhood. A graph is built on top of the trained tree via a carefully-designed edge weighting scheme that compactly integrates the constraints. A one-to-many graph search technique then serves as a diverse counterfactual generation scheme in low-entropy decision sub-spaces. CEnt is the *first model-agnostic* recourse method that does not pose differentiable requirements on the black-box model and satisfies immutability and semi-immutability constraints. It can also deal with categorical data and generate diverse counterfactuals that are attainable according to the underlying data distribution while allowing for user-defined feature costs. Our validation demonstrates the effectiveness of CEnt as it yields proximate counterfactuals while achieving high attainability and low latency and constraint violation rates. Our extension to imagery data and convolutional neural networks (CNNs) shows that CEnt can derive visual contrasts that are minimal and more useful than traditional explainability methods such as LIME [37]. Lastly, CEnt can highlight weaknesses of textual classifiers by deriving adversarial attacks.

Second, we present a Faithful Contrastive Fairness (FCF) as an individual fairness metric [38] that considers concentric neighbors around an individual and computes the corresponding contrastive cost. We define the contrastive cost to be the change required in the input features to modify a classification from an unfavorable outcome to a favorable one (or vice versa). A prediction is considered fair if these costs do not entail a change in the protected attribute or its dependents. More importantly, the neighborhood is derived based on a manifold-like distance metric computed by auto-encoders that account for data density and attainability. In the

same vein, 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* [39], [40], fairness frameworks and we propose tangible measures to quantify them. We empirically demonstrate how our metric satisfies faithfulness guarantees and detects bias cases missed by other methods.

Third, we extend contrastive fairness evaluation to textual settings where sensitive attributes are not explicitly manifested. We study the decision of a deep textual classifier by approximating its boundary in a local neighborhood. A model f is deemed fair if the computed decision boundary does not encode any sensitive information. We compute the sensitive attribute (SA) through Principal Component Analysis (PCA) or as a multivariate random variable with different realizations. Accordingly, we propose two measures inspired by geometrical analogies and mutual information. Within the Natural Language Processing (NLP) context, we extend the evaluation of contrastive faithfulness to textual data and we evaluate two state-of-the-art contrastive explanations models based on the proposed metrics. We study the effectiveness of our evaluation scheme on deep transformers and convolutional models in NLP using different SAs.

Finally, we devise a methodology to mitigate bias in classifiers and ensure our proposed contrastive fairness. The mitigation strategy can be incorporated during training to yield an inherently fair model. Alternatively, it can be used to neutralize a classifier in the latent space with no assumptions on the underlying classifier. More importantly, our mitigation scheme is not limited to tabular data; it can be extended to classifiers that operate on imagery and textual inputs. Our strategy is the first

augmentation technique that operates on the latent rather than the input space of classifiers. This alleviates the need for SA ontologies as a preparatory manual step. We demonstrate how our metric significantly reduces textual bias without compromising prediction quality. Additionally, we show that our mitigation yields an improvement in existing fairness metrics such as equality of opportunity and equalized odds.

The contribution of this dissertation can be summarized as follows.

- Objective I: We consider counterfactual explainability and we propose a novel contrastive explanation method.
  - We formulate a novel counterfactual explainability method, which minimizes the edit cost through graph search by employing a manifold-like distance metric learned by VAEs.
  - We extend our method to computer vision settings to derive visual contrasts between two confusing classes.
  - We extend CEnt to textual settings where we use the Bag-of-words representation to generate adversarial attacks on textual classifiers.
- Objective II: We consider fairness and we propose a novel *faithful* contrastive fairness (FCF) evaluation scheme.
  - We present FCF, the first fairness definition that can be extended to non-tabular data and that accounts for data distribution and attainability.
  - We devise a manifold-like distance computed by auto-encoders to derive similarities that reflect application-specific constraints.
  - We devise *faithfulness* evaluation of fairness measures and suggest corresponding quantification steps.

- Objective III: We extend the contrastive fairness evaluation and faithfulness guarantees to textual data
  - We quantify the sensitive attribute information in textual classifiers by applying PCA on the embeddings of predefined analogies.
  - We study the fairness of a classifier through metrics inspired by geometrical relationships and mutual information.
  - We extend faithfulness evaluation to textual data and we benchmark existing models accordingly.
- Objective IV: We suggest latent and contrastive bias mitigation methods to operate in pre- and post-hoc settings.
  - We devise an inherent bias mitigation technique to enforce contrastive fairness for classifiers during training.
  - Alternatively, we suggest the first augmentation technique that operates on the latent space to provide fairness guarantees at modest performance costs with little reliance on word ontologies.

The rest of this dissertation is organized as follows. First, we present the terminology that we follow in this work in Chapter 2 and we survey existing work on explainable AI, fairness evaluation, and bias mitigation in Chapter 3. Then, we examine the methodology of our objectives and their empirical results in Chapter 4, 5, 6 and 7. Lastly, we conclude in Chapter 8 and we highlight future research directions.

Chapter 2

Background

This chapter presents the terminology that we follow in this dissertation for explainable AI and the background for fairness.

2.1 Explainable AI: Terminology

Explainability methods are categorized according to their mode of application (the "when") and the knowledge type they are trying to interpret (the "what").

2.1.1 *The "when" of Explainability*

Based on the explanation mode, i.e., whether explainability is performed on pre-trained models or before building the architecture, ExAI methods can be categorized into:

Post-hoc interpretability These methods operate on pre-trained models with predefined architectures by analyzing how they process inputs before producing a decision. When no assumptions are made on the model, the interpretability method is model-agnostic. If the method entails a particular architecture, the method is model-specific.

Inherently interpretable models A less popular line of work develops models, from the ground up, that provide supporting evidence while processing input by modifying the underlying /architectures and learning strategies. The most famous examples of inherent interpretability are linear models and decision trees. The parameters in the former models quantify the feature importance and the decision in the latter models is a sequence of human-understandable if-then rules. In more complex models, generative models are used with other architectures to learn to generate fine-grained explanations while making classification decisions.

Despite their simplicity, inherently explainable models do not leverage state-of-the-art existing deep networks. For instance, retraining deep models such as the BERT-like family [41] and GPT models [42] with the explainability constraint can be computationally expensive. Retraining might not be feasible with strict privacy limitations or when training data is no longer available. On the other hand, explaining black box models in a post-hoc manner, rather than creating models that are interpretable in the first place, might perpetuate bad practices. Mainly, post-hoc explanations do not present perfect fidelity to the model being explained. They could be an inaccurate representation of the original model in the feature space. Therefore, the interpretability mode must be carefully chosen in high-stake areas such as criminology and law. Given that they harness the power of deep learning and the advancements in the training process and the use of transfer learning, post-hoc interpretability is gaining more interest within the ExAI community.

2.1.2 *The "what" of Explainability*

Explaining machine learning models occurs on three different levels: prediction (local and global), learning, and training. ExAI methods are ergo categorized, according to their explanation type, into four parts. **Local prediction** Local methods are

data-centric approaches that try to answer the question "why did the model predict y_1 on input x_1?" These methods provide explanations in terms of input features that are crucial for a particular output on a specific testing instance. Predictive models in critical areas can employ these methods to augment their predictions with supporting evidence to engender users' trust.

Global categorization Like local methods, global methods are also data-centric approaches. However, they try to answer the question "Why does the model predict an output y in general? What are the features that are crucial for the classification of this specific class/label?" They provide insights into the prominent features of the entire class in a trained classifier. The explanation is provided as a human-relatable concept that is common in all instances of a certain class/label. These methods can help AI practitioners debug their models and understand the contrast between predictions.

Learned knowledge and characterization of hidden features These methods are network-centric approaches that answer the question "What is this neuron trying to learn? or how can the neuron's activation be qualitatively explained?" They provide an interpretation of the internal state in terms of human-friendly concepts that apply across the entire dataset and go beyond per-sample features. In other words, they provide some matching between individual hidden units and a set of semantic concepts by studying how patterns are encoded in the hidden layers of a deep network. These methods are the main hope of rendering the black-box models less opaque by revealing their learned knowledge. They can also expose vulnerabilities, unintended correlations, and bias cases.

Learning dynamics These methods are network-centric approaches that study

the whole learning process and provide higher-level insights such as: "When is the class-specific information formed? What is the effect of freezing or the order in which the training instances are fed to the network?". In addition to this, they study how different training regimes affect the performance of deep neural networks.

2.2 Fairness Terminology

Let A be a set of protected attributes, sensitive attributes that each individual must not be discriminated against. Sensitive attributes include but are not limited to race, color, sex, sexual orientation, age, physical or mental disability, marital status, pregnancy, religion, political opinion, national extraction, and social origin. In each category, we are presented with a privileged and underprivileged group. While some attributes are binary (e.g. pregnancy), other attributes, such as race and religion, present a multi-label aspect.

We denote by X, the set of observable attributes and Y is the outcome. $\hat{Y}$ is the predictor that depends on A, X, and some other relevant latent attributes that are not observed. We note that Y might encode some historical or prejudicial biases. The goal is to study the fairness of the predictor $\hat{Y}$. For simplicity, we assume that $\hat{Y}$ has two outcomes: a favorable and an unfavorable one.

Individual fairness entails that two close individuals are treated alike by a predictor $\hat{Y}$. In other words, if they only differ by their protected attributes, and the privileged individual receives a favorable outcome, a similar treatment should be observed for the underprivileged one.

Group fairness considers the group, and studies the probability of assigning a favorable/unfavorable outcome for the privileged and underprivileged groups. More details on the different metrics are presented in Section 3.2.1.

Chapter 3

Related Work

Contrastive learning is proposed to enhance the performance of classification tasks by leveraging contrasts in the training data. Mainly, samples are contrasted against each other to teach the network about attributes that are common and distinct between classes. In other settings, contrastive learning aims at augmenting the dataset with versions of the same sample close to each other while forcing distinct embeddings for different samples [43]. Given its potential at enhancing performance, contrastive learning has recently become a dominant component in different learning paradigms such as supervised and self-supervised learning methods for computer vision and natural language processing.

For instance, contrastive learning is shown to be effective at enhancing clustering [44] debugging [21] and classification robustness through contrastive augmentation [45], [46]. Contrast sets are also extensively used in the evaluation of NLP models by exposing their vulnerabilities [21], [45], [47]–[51].

In the context of fairness, researchers augment their dataset with contrast examples where only the sensitive attribute is perturbed while maintaining the prediction outcome [35], [52]–[54]. These techniques are proved successful at neutralizing textual classifiers [36] and language models [52] while enhancing the robustness in most cases.

In what follows, we discuss how contrastive learning is used in explainable and fair AI. To this end, we report the related work to each of our three objectives. We start by surveying explainable AI techniques in their traditional and contrastive formats. Then, we discuss existing fairness measures and evaluation techniques in general settings, text, and imagery. We pay special attention to the distance measures used in such settings to introduce the faithfulness concept later. Finally, we survey existing bias mitigation techniques in different modalities.

3.1 Objective I: Contrastive Learning for Explainability

In what follows, we discuss state-of-the-art work on traditional methods in explainable AI. Then, we move to survey the counterfactual aspect of explainability.

3.1.1 *Explainable AI*

One of the first attempts at ExAI was carried out by [55] in 2010 by relying on local gradients. Their work addresses the question of why a black-box model predicted a particular label for a single instance and what input features contributed the most to a particular outcome. The estimation of the local gradient induced a quantification of the importance of a data point in altering the predicted label where higher gradients implied more important features. Based on the gradient concept, [56], [57] optimized the gradient-based approach and generalized its applications to cases where the computation of the gradients is problematic. Gradient-based methods are also congenial to imagery applications where individual pixels can be seen as features. In computer vision settings, saliency maps are the common term used to identify relevant regions in an image and to provide a type of quantification of the importance of a pixel in classification.

Gradient computation has also attracted researchers to develop efficient and fast methods for saliency computation. One of the earliest saliency methods was

proposed by [58]. Their method, which falls under sensitivity, computes the gradient of the class with respect to the image pixels and assumes that salient regions are at locations with high gradients based on the assumption that high gradient locations are important for the classification since their perturbation has a great effect on the output of the network. Later, many popular methods relied on the backpropagation of the gradient from the deepest layers of a network and its projection on the image to derive a gradient saliency map [18], [59], [60]. These techniques are widely used to explain visual recognition and object localization outcomes where relevant pixels are masked or highlighted (mostly with different intensities) as in [61], [62].

Instead of computing the gradient of the predictor with respect to an input feature, some approaches rely on perturbing the input and observing the impact on the predictor. Perturbation-based approaches construct explanations by analyzing the model's response to local changes in the input. [18] suggest performing local perturbations by masking portions of the input. Once some portions are occluded, a sensitivity analysis of the classifier output is performed to reveal the image parts that are important for the classification. Perturbation-based approaches are also appropriate in computer vision settings. For instance, the prediction difference analysis method of [63] samples within the pixel neighborhood to analyze the importance of an input feature. The relevance of a feature x is estimated by how the prediction of a class would change when all the features, except x, are used.

Another category in ExAI methods stems from neuroscience. To understand brain function, one of the most fundamental questions that the researcher tries to investigate is what types of stimuli excite neurons and drive them to fire? In the ExAI field, Activation Maximization (AM) methods address this question by focusing on the preferred stimuli that provoke a neuron in deep networks to fire strongly. AM techniques are heavily applied in imagery settings: given an input image $x \in \mathbb{R}^{H \times W \times C}$ and the parameters θ of a classifier, a neuron i_l in a layer l,

finding the image that maximizes the activation $a_{i_l}(\theta, x)$ is formulated by [64] asthe following optimization problem $x^* = \arg\max_x a_{i_l}(\theta, x)$. Gradient ascent algorithms are then used to solve the optimization at hand.

In natural language applications, some text-specific challenges hamper the application of general ExAI methods into NLP models due to the fusion of syntax and semantics in words, polysemy, and ambiguity. For instance, perturbation methods cannot be directly mapped to NLP due to the reliance of the latter models on embedding models that are opaque representations, as opposed to pixels or numerical values as we show in our work in [65]. Long-term dependencies, multi-lingual support, and learned stereotypes present additional challenges to ExAI techniques in NLP.

A great deal of ExAI techniques investigates the semantic and syntactical information learned by language models. In fact, several methods have been proposed to dissect the inner dynamics of transformers to better understand how they process input and why they do it so well. Such approaches can be the first building block in the process of making transformers trustworthy by rendering their inner workings understandable [66], [67]. They can also engender users' trust by explaining the knowledge learned by transformers and their parameters [68], [69] and by highlighting their limitations [70].

Due to their design, attention weights are relatively more interpretable than the conventional deep networks' parameters. Visualizing the inner weights and hidden representations of transformers can render the predictions more explainable [71]–[73]. However, solely, attention weights are not able to provide the full transparency that responsible AI entails where further processing is needed when the task at hand is not a simple classification but a more complex task such as translation, question answering, and natural language inference [14], [74].

3.1.2 *Counterfactual Explainability*

While the definition of parameter sensitivity differ with each method, the majority of these explanation methods operate on features that are *present* in the input, even if they may result in explaining features that are negatively contributing to the decision of the ML model [75]. One can alternatively explain a model decision by the necessary and/or sufficient condition that should be present or absent or even changed to justify or a particular decision by the model. Counterfactual explainability methods look for proximate input that can alter the model prediction from y_1 to y_2. Formally, assuming a predictor, potentially non-linear, $f : \mathcal{X} \mapsto \mathcal{Y}$, an instance $\boldsymbol{x^i} \in \mathcal{X}$ such that $f(\boldsymbol{x^i}) = y_{\text{fact}}$ and a foil class y_{foil}, a counterfactual $\boldsymbol{x}^i_{cf} \in \mathcal{X}$ can be computed by:

$$\underset{\boldsymbol{x}_{cf}}{\arg\min} \quad d(\boldsymbol{x}^i_{cf}, \boldsymbol{x^i}) \tag{3.1}$$

$$\text{subject to} \quad f(\boldsymbol{x}^i_{cf}) = y_{\text{foil}} \tag{3.2}$$

where $d(.)$ is a distance metric. This optimization can be also perceived as

$$\underset{\boldsymbol{x}_{cf}}{\arg\min} \quad L(f(\boldsymbol{x}^i_{cf}), y_{\text{foil}}) + \lambda d(\boldsymbol{x}^i_{cf}, \boldsymbol{x^i}) \tag{3.3}$$

in the Lagrangian notation, with $l(.)$ denoting a loss function and $\lambda > 0$ is a regularization factor that balances the minimal edit distance (through minimizing the distance between the factual and the counterfactual) and the success rate (by finding a counterfactual that successfully changes the model's decision).

Depending on the method, restrictions and assumptions might complement the above definition. For instance, model-specific recourse methods [24] put assumptions on f whereas model-agnostic ones assume a black-box architectures [76], [77].

Miller et al. [78] introduced contrastive explanations by relying on foundations in

philosophy and cognitive science as two types of contrastive why-questions: alternative why–questions, addressing the "rather than" part, and congruent why–questions, addressing the "but" part. Very recently, Dhurandhar et al. [75] argued that such forms can be found in many human-critical domains such as medicine and criminology. Thus, Dhurandhar et al. were the first researchers to propose a novel method that, given an input, finds the contrastive perturbations—minimal changes that are required to change a particular decision for *any* black-box deep model. This is achieved by solving an optimization problem searching for the minimal sufficient condition that needs to be present (in the case of pertinent positive explanation) or absent (in the case of pertinent positive explanation) in the input to change its classification. Their approach is validated on MNIST, a large procurement fraud and a brain activity strength datasets.

Prior to that, Ribeiro et al. [79] searched for the sufficient conditions to justify classification decisions for a particular category or class. Consequently, instead of computing the perturbation needed in a specific instance for the output to be perturbed as in [75], [79] compute the feature values that can imply the whole class as global rules (or *anchors*). The approach is validated on tabular, textual and imagery datasets in a set of classification, structured prediction and text generation tasks.

Later, [80] generalize the contrast suggested by [75] as a contrast between the produced output, *fact*, and any arbitrary other class to the contrast between the *fact*, and a specific output of interest, the *foil*. This is achieved by training a decision tree centred around a particular point of interest to learn the contrast between the fact and the foil. Once the tree is computed and the fact leaf is located, search algorithms are employed to search for the nearest foil leaf which results in a set of rules that represent the contrastive explanation of the model's decision. Their approach is validated on three tabular benchmarking datasets: Iris, PIMA Indians

Diabetes and the Cleveland Heart Disease datasets [81].

Instead of approximating models by local trees as in [82], [83] computed contrastive explanations by relying on the SHAP method introduced by [20] as a unified approach to interpret machine learning models predictions in a model-agnostic manner. Their pipeline is tested and discussed on the IRIS, wine quality and the mobile features dataset. Dhurandhar et al. [84] also considered model-agnostic explanations, for structured datasets, having both real and categorical features. The validation on diverse datasets proved the outperformance of the counterfactuals over traditional explanations provided by [85]. The effectivess of these counterfactual explanations lead to their application on Reinforcement Learning (RL) environments in [82] where the contrast of interest is designed to be between the consequences of the user's query-derived policy and the optimal policy learned by an RL agent.

In summary, researchers are recently showing a growing interest in contrastive methods that resemble human argumentation and that are showing their effectiveness in different domains. However, such methods are still in their vanilla versions. For instance, different weighting techniques suggested by [22] can be applied to the work of [80] and advanced search algorithm have the potential of enhancing their contrastive explanations. Moreover, the majority of these methods do not test on a set of well defined benchmarking data, and are bounded to tabular or structured datasets. In this work, we intend of extending the work of [80] by considering edge weight in advanced tree-building strategies while validating on a new set of tasks, including textual data if time permits. Table 3.1 shows a summary of existing methods that are relevant to our work aggregated by their formulation of the contrastive explanation, the data type they are evaluated on, the method used to compute the contrastive explanation and whether they are model-agnostic or not.

	Model-agnostic	Immutability	Categorical	Semi-immutability	Diversity	Generated
CEM [28]	✗	✗	✗	✗	✗	✓
AR [24]	✗	✓	Binary	✓	✗	✓
[23]	✗	✗	Binary	✗	✗	✓
GS [86]	✓	✓	Binary	✗	can be extended	✓
REVISE [87]	✗	Binary	Binary	✗	✗	✓
CLUE	✗	✗	✓	✗	✗	✓
FACE [76]	✓	Binary	Binary	can be extended	✗	✗
DiCE [27]	✗	✓	Binary	✓(post-hoc)	✓	✓
CRUDS [88]	✓	✓	✗	✓	✓	✓
CEnt	✓	✓	✓	✓	✓	✓

Table 3.1: Summary of existing contrastive explanation methods based on their underlying assumptions (gradient-based approaches are not model-agnostic) and on whether they handle immutable, semi-immutable, and categorical features. We also highlight methods that can generate diverse counterfactual examples (CEs) and those whose CEs are generated or selected from the training set.

3.2 Objective II: Faithful Fairness Evaluation (FCF)

We dedicate the next part to surveying existing metrics for fairness evaluation and we discuss their limitations. Then, we discuss the distance measures used in the evaluation schemes before introducing the manifold-like distance scheme.

3.2.1 *Existing Fairness Definitions*

Many definitions for fairness exist in the literature [16], [29]. In what follows, we summarize some of the causality-based and statistical ones.

Definition 1 (Fairness Through Unawareness, FTU)**.** *A predictor is fair if the protected attributes are not explicitly used in the decision-making process. Formally,* $\hat{Y}$ *is fair if* $\hat{Y}(X, A) = \hat{Y}(X)$.

For instance, any mapping $\hat{Y} : X \mapsto Y$ that excludes A from X satisfies this definition. Despite its simplicity, fairness through unawareness does not account for the case where the sensitive attribute is not used in the decision-making process but its dependents are. While the literature [31] highlights the aforementioned shortcoming; we argue that an unpretentious exclusion of the features dependent on A

can resolve the issue. Their exclusion, however, might compromise the performance of $\hat{Y}$ highlighting therefore the trade-off between fairness and accuracy.

Definition 2 (Individual Fairness, IF)**.** *A predictor is fair if for similar individuals the prediction is the same. Formally, $\hat{Y}$ is fair if for individuals i and j where $d(i,j)$ is small, their prediction $\hat{Y}(X^{(i)}, A^{(i)}) \sim \hat{Y}(X^{(j)}, A^{(j)})$ for a distance metric $d(.,.)$[38].*

According to [38], $d(.,.)$ needs careful selection and understanding of the application domain. We add to this shortcoming the following empirical considerations. First, in the case of multiple protected attributes, or when the set of features dependent on A, $DEP(A)$, is considered, the similarity measure between i and j is not easily defined. Furthermore, the notation does not take into consideration the $\frac{\partial \hat{Y}}{\partial A^{(i)}}$ and the implementation needs to exhaust the neighborhood of an individual according to $d(.,.)$.

Definition 3 (Demographic/Statistical Parity, DP)**.** *A predictor is fair if both privileged and underprivileged groups have the same probability of being assigned to the favorable outcome c. Formally, $P(\hat{Y} = c_{favorable}|A = 1) = P(\hat{Y} = c_{favorable}|A = 0)$.*

This definition ignores any possible correlation between Y and A. Consequently, fairness with respect to demographic parity requires that, for all groups, the same ratios are selected for each class c. This might compromise the model's performance as well.

Definition 4 (Equalized Odds)**.** *A predictor is fair if the prediction is independent of A conditioned on Y which leads to the definition of the equality of opportunity.*

Definition 5 (Equality of Opportunity)**.** *A predictor is fair if the positive prediction is independent of $A\|Y$.*

Trying to satisfy the last two fairness definitions can increase the gap between privileged and underprivileged groups. For example: trying to satisfy the same

acceptance rate for groups A and B in a hiring algorithm will lead to accepting more A if the number of qualified candidates from group A was higher. Consequently, A might have a higher income and can thus afford better education for their children who will more likely get well-paid jobs increasing thus the gap between A and B. Similar definitions include but are not limited to predictive parity, predictive equality, and disparate impact.

In [31], the authors argue that it is not immediately clear how to tackle historical bias in the aforementioned fairness definitions. Moreover, the dependency between attributes should be taken into consideration, even if this necessitates strong assumptions. To remedy that, they study fairness from a causal perspective by introducing counterfactual fairness.

Definition 6 (Counterfactual Fairness)**.** *A predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$, $P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)$ for all y and any value a' attainable by A with U being the set of relevant latent attributes which are not observed.*

This definition enforces that the distribution over possible predictions remains unchanged when protected attributes are causally different. However, its generalization to datasets where the protected attributes are implied (such as images) is impractical. In such cases, the alteration of the attributes is not as easy as a simple modification of a tabular dataset. Additionally, the authors consider permutations of protected attributes and their (causally) dependents which can have two limitations. First, in the case where multiple attributes are correlated, or causally dependent on A, exhausting different combinations can thus be computationally prohibitive. Second, the case where alteration of A, coupled with other unrelated attributes can change the model's prediction.

To illustrate that, we consider a hiring algorithm that takes as input a list of attributes along with gender (sensitive) and GPA. We assume that a candidate

woman with 3.3 GPA is not selected by the algorithm $y(x, G = W, GPA = 3.7) = 0$. We consider 4 counterfactuals and their predictions as follows: $y(x, G = M, GPA = 3.7) = 0$, $y(x, G = M, GPA = 3.8) = 1$, $y(x, G = W, GPA = 3.8) = 0$ and $y(x, G = W, GPA = 3.93) = 1$. This is clear discrimination against women: with the same qualifications, a woman needs a GPA of 3.93 to be hired whereas 3.8 is sufficient for a man. However, [31] consider this as *counterfactually* fair since altering the protected attribute only does not change the prediction ($y(x, G = M, GPA = 3.7) = 0$). In this example, gender is not a direct cause for hiring; it is instead a legacy that can increase the chance of being hired.

3.2.2 *Distance Measures and Faithfulness*

Let x_i and x_j be two observable inputs. The distance between x_i and x_j is usually computed as the p-norm distance: $d(x_i, x_j) = \left(\sum_{l=0}^{d} |x_i^{(l)} - x_j^{(l)}|^p \right)^{1/p}$. The literature has formed a consensus on the use of l_0, l_1 or any normalized convex combination thereof [23], [24], [89], [90].

We often face a combination of continuous, ordinal, and nominal inputs concurrently in domains such as healthcare or the legal domain. For this heterogeneous type of data, it might be not meaningful in measuring the distance as described above. For instance, consider the aforementioned hiring example. Is candidate A (G=W, GPA=3.7) closer to B (x, G=W, GPA=3.8) or C (x, G=M, GPA=3.7)?

Some methods leave the selection of appropriate distance measures to application-specific experts [23], [91]–[94] which can be considerably subjective. Additionally, task-specific definitions have their own ethical issues hindering the practicality of individual fairness [95]. To alleviate the reliance on a distance metric, Hu and Rangwala [96] proposed a metric-free definition of individual fairness through cooperative contextual bandits.

Recently, there have been some approaches [97]–[99] that non-linearly model

latent spaces using adversarial learning [9] techniques. Other approaches [100]–[102] use Variational Auto-Encoders (VAEs) [103] to regularise the latent space by removing low-density regions. To the best of our knowledge, these distance measures are not utilized by fairness definitions yet. Fig. 3.1 shows the importance of manifold-like (geodesic) distance in neighborhood selection. Two points can be deemed neighbors based on Euclidean distance (L2-norms). However, according to the manifold distance, that reflects true distance, those points might not belong to the same neighborhood. Consequently, in the fairness framework, it might not be effective to compare an observation to a neighbor computed according to the Euclidean distance.

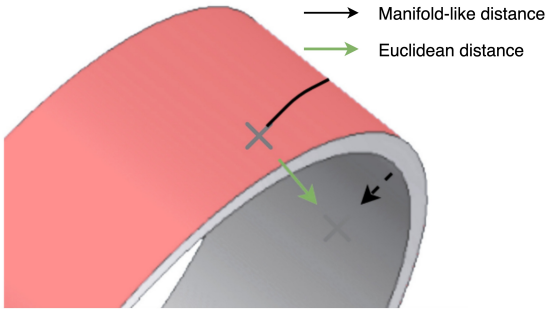


Figure 3.1: Importance of manifold-like distance metric in neighborhood sampling

3.3 Objective II: Fairness Evaluation in Textual Settings

In this section, we discuss how existing fairness metrics can be extended to textual settings. Additionally, we survey existing work on evaluating textual contrasts.

3.3.1 *Fairness Evaluation in Textual Classifiers*

Researchers extend the definitions presented in Section 3.2.1 to textual settings. For instance, [48] extends CF's concept to textual applications where sensitive words are perturbed in the text. To date, the prevailing methods within prior research relied on input perturbations [35], [36], [52]–[54] to detect discrimination with respect to

gender mostly.

While successful at evaluating fairness in general settings, fairness definitions share the same limitations in text. They embody the presumption of the availability of membership labels and that sensitive attributes are explicitly revealed in the text. For FTU, words that contain discriminatory information are not obvious to detect specifically with correlations. For instance, *race* can still be leaked by *neighborhood* information even if specific words were substituted. Textual modality further amplifies FTU's failure at achieving fairness with grammatical gender (e.g. `actress` and `actor` carry gender information). Furthermore, language models encode stereotypes in their context rendering gender unawareness ineffective at avoiding stereotypes. In fact, BERT [6], when trained to predict gender from biographies after removing gender-specific pronouns, can still infer *male* from the biography of a surgeon for example (See Figure 3.2). De-biasing embeddings [104] is an FTU application to text that does not go unchallenged due to grammatical gender and equivocal word correlations.

Finally, counterfactual fairness of [31] requires locating words with sensitive attribute info which might be a single point of failure. For example, for *religion*, *Christmas*, *kippah*, *Mecca* encode sensitive information which is not evident. Second, substitution might not be an adequate counterfactual generation method; insertion, deletion and substitution of word segments should be considered as well. Finally, word substitution might lead to grammatical inconsistencies. It may even be impractical (e.g. substituting religious artifacts such as *Christmas tree* or *Diwali lights* [36]). It might also not be sufficient to detect bias (detecting ethnical bias in `there are a lot of actors trying to be funny`) requires inserting a word such as `Chinese` or `black`) which makes the search space on the perturbed inputs extremely large [53].

In this work, we do not rely on perturbations and we operate on latent repre-

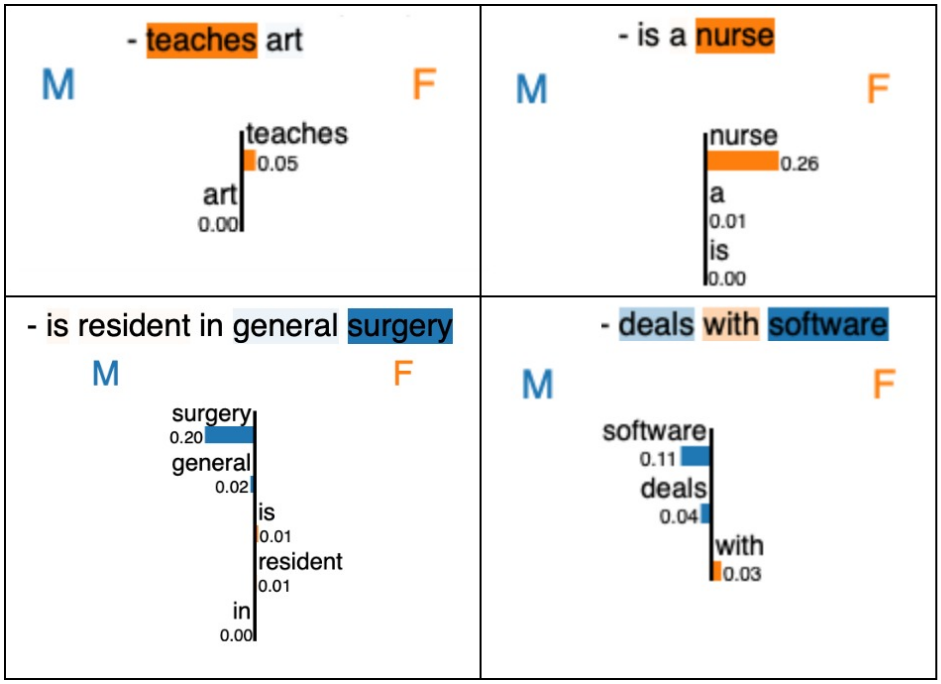


Figure 3.2: LIME explanations of stereotypes on gender prediction. Pronouns revealing gender are hidden. Orange (blue resp.) indicates female (male resp.).

sentations to overcome the aforementioned challenges in evaluation and mitigation. Our formulation does not require explicit reference to sensitive data which addresses the limitations of previous fairness definitions.

3.3.2 *Evaluation of Contrastive Examples in NLP*

The evaluation of textual contrastive sets is addressed from qualitative and quantitative perspectives.

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 3.1. In NLP, these conditions are referred to as *minimal distance* and label-flip score respectively. The proximity between $\boldsymbol{x}_{cf}$ and $\boldsymbol{x}$ is measured by word-level Levenshtein distance [105] 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 [106] 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 [45], [107].

An additional requirement for counterfactuals is the diversity of the generated explanations. Inspired by the Self-BLEU metric of [108], diversity can be measured through the Self-BLEU or Self-BERT [109] metric between the generated counterfactual samples. A higher BLEU score implies similar counterfactuals and thus a less diverse contrastive generation.

Other requirements that are tailored to natural language are (1) fluency through grammatical correctness and semantic meaningfulness, and (2) content preservation. Fluency operationalizes "probable" contrastive texts that are not the result of a coincidence. [110] highlights a strong human preference for counterfactuals that are proximate to the original text but that are highly probable based on the original distribution and are not caused by rare events. Fluency thus measures the similarity between the distributions of the counterfactuals and the original data. 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 [45], [111], [112]. Content preservation can be inferred by latent embedding representations as the cosine similarity between the embeddings of $\boldsymbol{x}_{cf}$ and $\boldsymbol{x}$.

Qualitative Evaluation

In the social aspects of AI, user studies are ubiquitous in evaluating explainable and fair AI models [113]–[115]. 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, sat-

isfaction, and understandability mainly. Instead 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.

3.4 Objective III: Bias Mitigation

Bias mitigation in machine learning classification models is usually performed as a pre-, in- or post-processing step [116]. In-processing methods remove bias from the training data, whereas in-processing methods are applied during training and post-processing methods are performed on trained models in a post-hoc manner.

Removing bias before training can be done through re-labeling and correction of discrimination cases [117], [118]. Another famous approach is data augmentation and re-weighing of training samples [119]–[122]. In image classification, this is by generating realistic images from under-represented groups [123] to enhance fairness. In textual settings, analog sentences, with different sensitive information, are augmented with the data set to ensure a similar treatment [36]. Additionally, representation learning is employed as a preprocessing step to learn a transformation of training data such that bias is reduced. This manifests itself mainly in textual settings, where different approaches have been developed to neutralize word embeddings [52], [104], [124], [125].

To enforce a fairness constraint while training, researchers proposed the use of regularization and constraints [126], [127] and to give importance to the correct or unbiased features [128]. Another line of work exploits adversarial learning to ensure fairness guarantees. The classification model is trained to minimize a loss function and the adversary is trained to exploit fairness issues. In a min-max game, both models then compete against each other to improve the performance while minimizing bias. Examples of adversarial learning for bias mitigation include, but

are not limited to, [129]–[132].

Post-processing techniques focus on applying correction steps to classifiers that predictors. This is mainly achieved by solving an optimization problem under consideration of fairness loss terms (with respect to equalized odds) [133]. A similar technique is applied to causal models [134] the impact of the sensitive attribute on the prediction is corrected to ensure the counterfactual fairness of [31]. [135] relabeled individuals who are likely to receive biased results according to an individual bias detector.

Chapter 4

Objective I: Entropy-based Contrastive Explanations

Existing counterfactual explainability methods surveyed in Section 3.1 often employ causal graphs, gradient descent, discriminative, and evolutionary algorithms to generate contrastive examples (CEs) while satisfying feasibility constraints [24]–[26]. Apart from being able to contrast the outputs of various models, the most wanted desiderata of counterfactuals are plausibility, attainability, and diversity. While contrasting the output is successfully achieved by all methods, other requirements partake in a trade-off and are rarely simultaneously satisfied. For instance, some methods violate constraints [23], and others do not always produce attainable counterfactuals [27] or produce a unique CE based on a proximity measure [28].

However, the daunting acquisition of causal graphs and the unavailability of user-defined similarity measures limit the adoption of such techniques in practice [77]. Furthermore, gradient-based methods are sensitive to the classification boundary and the geometry of the underlying data distribution [88]. Operating on features that are *present* in the input even if their perturbation might yield explanations that contribute negatively to the decision-making process [75]. More importantly, existing methods prevent downstream users from exploring alternatives and specifying

the cost of alterations in an ad-hoc manner.

In this work, we address the shortcomings of existing methods and design a **C**ontrastive **Ent**ropy-based Explainability method, CEnt, under feasibility, immutability, and semi-immutability constraints while satisfying proximity and user-defined costs. CEnt can also deal with categorical data and generates diverse counterfactuals that are attainable according to the underlying data distribution while allowing for user-defined feature costs. Given an observation x, CEnt samples k local neighbors of x based on manifold-like distance approximated by Variational Auto-Encoders (VAEs). Then, CEnt approximates a black-box machine learning model by a decision tree in the local neighborhood. A graph is built on top of the trained tree via a carefully-designed edge weighting scheme that compactly integrates the constraints. A one-to-many graph search technique serves as a diverse counterfactual generation scheme in low-entropy decision sub-spaces.

This chapter presents the problem statement of our first dissertation objective in Section 4.1 before describing our methodology and studying its complexity in 4.2. We then validate CEnt on different data types and model architectures in Section 4.3 before highlighting the limitations and future directions in Section 4.4.

4.1 Problem Statement

We consider a multi-class black-box classifier $f : \mathcal{X} \mapsto [0,1]^{|C|}$ with $f(x)$ being a c-dimensional vector specifying the probability of x belonging to each class in C. We consider an input $x, f(x) = y_{\texttt{fact}}$ for which we would like to derive a close CE $x', f(x') = y_{\texttt{contrast}}$. We define $g : \mathcal{Z} \mapsto C$ to be an entropy-based approximation of f. Given a proximity measure π and an edit distance δ_g, the contrastive $x' \in \mathcal{X}$ is obtained by minimizing $\delta_g(x, x')$, and the approximation loss calculated on local neighbors of x, $\mathcal{L}_\pi(f, g)$ while imposing a regularization component. An additional constraint is for the model prediction on x' to be the desired contrast class. The

problem can thus be formulated as:

$$\arg\min_{x'} \quad \mathcal{L}_{\pi_x}(f, g) + \lambda_1 R(g) + \lambda_2 \delta_g(x, x') \tag{4.1}$$

$$\text{subject to} \quad f(x') = y_{\texttt{contrast}} \tag{4.2}$$

, with λ_1 and λ_2 are regularization parameters on the complexity of g and the proximity measure respectively.

We refer to the approximation loss $\mathcal{L}_{\pi_x}(f, g)$ as locality-aware fidelity loss. We imply that a model g that minimizes the loss of fidelity should produce results similar to f in the local neighborhood $\tilde{\pi}_x$. Assuming that a model g is a *faithful* approximation of f, the constraint can be replaced by $g(x') = y_{\texttt{contrast}}$.

In this chapter, we showcase how we minimize the objective above and we empirically validate our approach.

4.2 Methodology

The model-agnosticism requirement of our approach thwarts any assumptions on f, thus, any gradient-based solution. Alternatively, we force the constraint by reducing our search space to nodes in the Decision Tree (DT) corresponding to $g(x') = y_{\texttt{contrast}}$ that minimizes the locality-aware fidelity loss $\mathcal{L}_{\pi_x}(f, g)$. We then minimize the edit distance through our one-to-many shortest path problem based on contrast boundaries learned by g. Consequently, we encourage feature changes with low entropy (i.e. high info gain) that can alter decisions. Our methodology is visualized in Figure 4.1.

4.2.1 *Minimizing Locality-aware Fidelity Loss*

We *locally* approximate the behavior of f with a DT. DTs are favored given their ability to provide a range of CEs rather than single points in the counterfactual

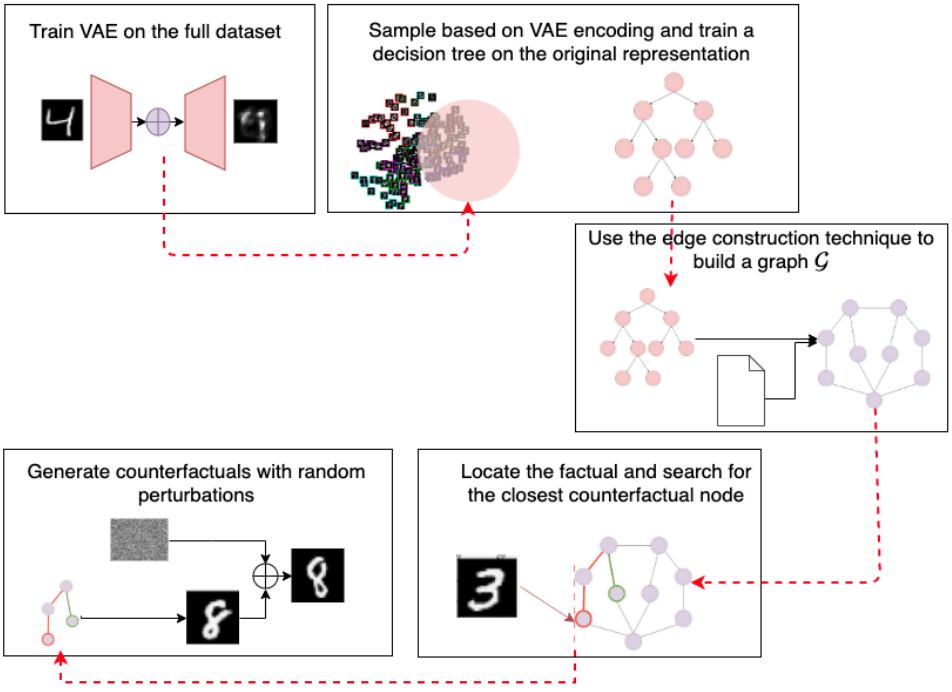


Figure 4.1: Overview of CEnt

world. Additionally, simple g models are desired as they are highly interpretable but might not yield good approximations. We model this trade-off by minimizing $\mathcal{L}_\pi(f, g)$ while maintaining the low complexity of g through the regularizer $R(g)$. Lastly, to satisfy the immutability requirement of some features such as gender, we remove such features from the input.

Sampling in the local neighborhood

Existing work computes distances between samples based on $L - p$ norms, most akin in $p = 0$ (edit distance) or $p = 2$ (edit L2 cost) or on domain experts to elicit the appropriate distance function [91], [136]. We employ a manifold-like distance to approximate actual distances while reflecting attainability. Such non-Euclidean distance is also suitable for categorical and non-tabular data. We mainly utilize VAEs which demonstrated significant performance gains in approximating manifold-like distances. Mainly, VAEs learn a new geometry, potentially denser, of the intervention space while encoding correlations, feasibility, and the plausibility of a CE

occurring.

Hence, given the input x, we learn its latent representation z in a self-supervised manner by defining a latent model $p(x) = \int p(x|z)p(z)dz$, an encoder $m(.)$ with its parameters Φ and distribution $q_\Phi(z|x)$, and a decoder $g(.)$ parameterized by Θ with a likelihood of $p_\Theta(x|z)$. Φ and Θ are represented by potentially non-linear functions. We then define $p_\mathcal{D}(x)$ as the empirical distribution of the data. Under these circumstances, the evidence lower bound (ELBO) [137] can be used to compute the intractable integral above as:

$$\mathbb{E}_{p_\mathcal{D}}(x)\Big[\log p_\Theta(x)\Big] \geq \mathbb{E}_{p_\mathcal{D}}(x)\Big[\mathbb{E}_{q\Phi(z|x)}\big[\log p_\Theta(x|z)\big] - \mathbb{KL}\Big(q_\Phi(z|x)||p(z)\Big)\Big] \quad (4.3)$$

where $\mathbb{KL}$ is the relative entropy or the Kullback–Leibler divergence [138]. Moreover, $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\Theta(\boldsymbol{x}|\boldsymbol{z})$ are assumed to be Gaussian. In the case of binary attributes, the decoder can be assumed to be Bernoulli. Once computed, the encoding z will be utilized to compute proximity as $\pi(x, x') = ||z - z'||_2$.

4.2.2 *Minimizing Counterfactual Cost Through Graph Search*

g gives a family of counterfactuals inferred from every leaf node labeled $y_{\texttt{contrast}}$. The goal is to search for the most proximate counterfactual. The search path $y_{\texttt{fact}} \rightsquigarrow y_{\texttt{contrast}}$ can be directly translated into a CE. To this end, we construct the directed weighted graph $\mathcal{G} = (\mathcal{V}, E)$ with $\mathcal{V}$ constructed from g nodes where $v \in \mathcal{V}$ can be a leaf node (class label) or an internal node (decision). The goal is to reach $y_{\texttt{contrast}}$ from $y_{\texttt{fact}}$ through the path reflecting the minimal edit.

Edge construction

We consider the edge $e_{ij} \in E$ connecting decision node i to decision/label node j. $e_{i,j}$ represents a decision $f_i \bigoplus v$, where f_i is the feature in the node i, $\bigoplus$ is an operator and v is a threshold value or a category. $e_{i,j}$ is proportional to the edit cost of f_i. We assume similar costs of edges e_{ij} except for the following cases:

- Custom cost function where the user specifies an edit cost c_{edit} of f_i (e.g. the cost of changing a job is twice that of relocating). In this case, all edges e_{ik} inherit c_{edit}.

- Semi-immutability where editing feature i is only possible in a particular direction (e.g *has_degree* cannot be set to *false* when it is *true*). In view of this, we assign an infinite weight to the corresponding edge.

Graph search

The vertices $v \in \mathcal{V}$ correspond to one of the following labels: (1) `fact`, (2) `contrast`, and (3) internal decision node. We identify, u_{start} the *unique* start node from category (1) that corresponds to x. Then, we execute a one-to-many shortest path problem on $\mathcal{G}$ from u_{start} to nodes in category (3). Once the search resumes, the result will be in the form of feasible rules $f_i \bigoplus v$. For practitioners who are interested in the CE instead of the contrastive path, we derive x' as follows. We consider f to be a feature in x. If f is not part of the contrastive path, its value is kept intact in x'. Otherwise, it is altered according to the $f_i \bigoplus v$ with a margin $\sim \mathcal{N}(0, \frac{\sigma_i}{m})$ with m is a tunable parameter and σ_i is the standard deviation of f_i. For categorical values, no random perturbations are applied.

4.2.3 *Complexity*

Lastly, we study the complexity of CEnt that is comprised of three main components: local sampling, DT training, and graph search. The former two components are extensively studied for optimizations in the literature through vector quantization [139], pre-pruning and ensembling. We focus our study on the third component which constitutes the main building block of CEnt. The problem can be cast into a single source with non-negative edge weights and no cycles; thus Dijkstra's algorithm is a suitable infrastructure. With a Fibonacci, instead of a binary, heap, the complexity can be optimized to $\mathcal{O}(E + V \log V)$ [140]. Furthermore, the constrained

construction of $\mathcal{G}$ gives rise to the following guarantees:

- $|V| <<< 2^{\texttt{max\_depth}}$ is controlled by (1) the size of input that affects the *max_depth* parameter of the DT and (2) the pruning techniques that avoid over-fitting.
- $|E| \leq |V| - 1$. Equality holds only when no semi-immutability constraints are imposed.

Accordingly, the complexity becomes $\mathcal{O}(V \log V)$ with a bounded $|V|$ that follows from small max_depth parameters.

4.3 Results

In this section, we validate CEnt on a variety of datasets. We demonstrate its extension to imagery data and a special use case for detecting vulnerabilities of textual classifiers.

4.3.1 *Experimental Setup*

We implement CEnt within CARLA framework [141] which also provides an implementation of existing recourse techniques [1]. Experiments are run on 2 cores of Intel(R) Xeon(R) with 12GB RAM. We train 2 models on the numerical datasets: a logistic regression (LR) model and a neural network (NN). NN consists of 2 layers with 13 and 4 neurons activated via *relu* and trained using a weighted binary cross-entropy loss function through gradient descent with root mean squared propagation.

Numerical Datasets

Four numerical datasets [141] are used in this work. The Adult dataset is used to predict whether an individual has an income >= 50K USD/year and consists of 48,842 instances and 14 attributes with *age*, *sex* and *race* set as immutable.

[1] https://github.com/carla-recourse/CARLA

COMPAS consists of information about more than 10,000 criminal defendants and is used by the jurisdiction to score the re-offending likelihood. The immutable features of COMPAS are *sex* and *race*. The Credit dataset consists of 150,000 attributes and 11 features to predict the possibility of financial distress within the next two years, *age* being the only immutable feature. Finally, the HELOC dataset consists of 21 attributes that describe anonymized information about home equity line of credit applications made by real homeowners. HELOC has 9871 instances used to predict whether the homeowner qualifies for a line of credit or not based on 21 features with no immutability constraints.

CEnt Settings

We train a VAE with batch normalization for 10 epochs with a learning rate of 0.001 and a dropout rate of 0.2. The weight used in the KL divergence is 2.5×10^4. The number of hidden layers and neurons in our VAE is adaptive to the input size and is selected according to the best validation loss. For adult data, we used one layer of 25 neurons and a bottleneck of size 8. For Credit and COMPAS, we use a layer of 16 neurons and a bottleneck of size 7 whereas, for HELOC, we utilize 2 hidden layers with 25 and 16 neurons and a bottleneck of size 12. For image datasets, we employ 3 layers of 500, 250, and 32 neurons. Sampling k neighbors is efficiently achieved using [142][2] with k=1000 equally distributed among the `fact` and `contrast` classes. We split our data into 80% training used to train f and 20% testing used to test CEnt and other contrastive methods. We set the *max_search* parameter to 50, i.e. we try at most 50 diverse CEs, if none flips the prediction, we claim failure to produce a counterfactual, and we return the one tried at last.

Metrics

We employ 9 metrics to assess the following aspects in the contrastive search.

- *Fidelity.* We evaluate the local performance of the DT, g, in approximating f

[2]https://github.com/lmcinnes/pynndescent

through the accuracy of the model with respect to the predictions of f'. We also compute the rate of semi-immutability constraint violations (*age* cannot decrease in the CE) and immutability violations. Fidelity is reflected by higher accuracies and lower violation rates.

- *Proximity.* We evaluate how close the derived CE x' is to its original counterpart x. $l0$-cost computes the number of feature changes between x and x'. $l2$-norm reflects the Euclidean distance between x and x' as $\sqrt{\sum_i (x_i - x'_i)^2}$. We also compute π, the $l2$-distance on the VAE encodings to reflect the manifold-like distance as $\sqrt{\sum_i (z_i - z'_i)^2}$. Finally, we compute redundancy, as in [141], to evaluate how many of the proposed feature contrasts were not necessary. This is achieved by successive flipping operations of values in x' into x and inspecting whether the label would flip back. Lower distances and redundancy scores are favored.

- *Flip rate.* We test the ability of x' in changing the prediction (a.k.a success rate). Higher scores are favored.

- *Latency.* We measure the time needed to derive a counterfactual in seconds. In the methods that require VAE encodings, training VAEs is excluded from latency calculation but obtaining the encodings is included.

- *Agreement.* We finally measure the agreement between x' and its neighbors by computing the **yNN** score as in [141] with $k = 5$. A score ≈ 1 implies that the neighborhood consists of points with the same predicted label CE x'; thus an attainable CE.

4.3.2 *CEnt on Numerical Datasets*

We randomly sample 100 instances from the testing data equally distributed among positive and negative labels, and we test CEnt on both models and all 4 datasets.

We compare CEnt against CCH-VAE [40], CEM [28], GS [86], CLUE [143], FACE [76], DiCE [27] and CRUDS [88].

	LR		NN	
	f	g	f	g
Adult	84	95	84	94
COMPAS	84	97	81	96
Credit	93	98	93	99
HELOC	73	87	72	89

Table 4.1: Accuracies on the original model f and the DT g

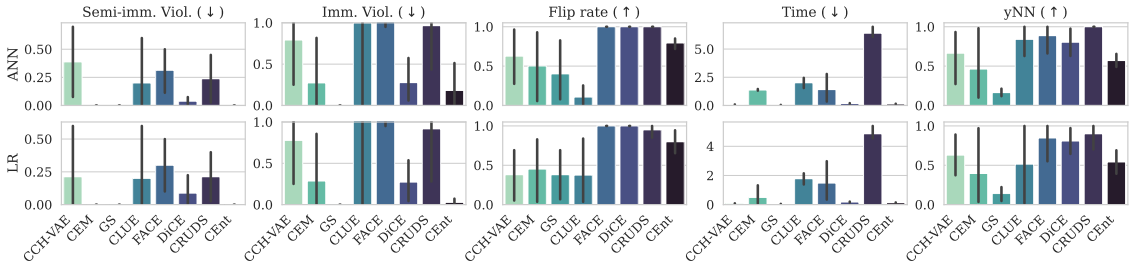


Figure 4.2: CEnt results averaged on four numerical datasets

Fidelity with respect to f is reflected in high g scores, implying an accurate approximation, as shown in Table 4.1. The percentage of violations, flip rate, latency, and agreement are reported in Figure 4.2 in the LR and ANN models averaged across datasets. CEnt consistently respects immutability and semi-immutability constraints that are significantly violated with methods such as FACE and CLUE. The results also show that CEnt can derive CEs in < 1 sec that can successfully change the prediction with a $\sim 90\%$ probability. Finally, the yNN scores surpass CCH-VAE, CEM, and GS but are lower than FACE, DiCE, and CRUDS which shows competitive attainability scores.

The distribution of the average proximity metrics per model is shown in Figures 4.3 and 4.4 across all datasets. CEnt achieves significantly low edit distances ($l0$) in most cases. Similarly, $l2-$ and VAE-distances are small except in some cases where methods such as CEM and DiCE can derive closer CEs. It is worth men-

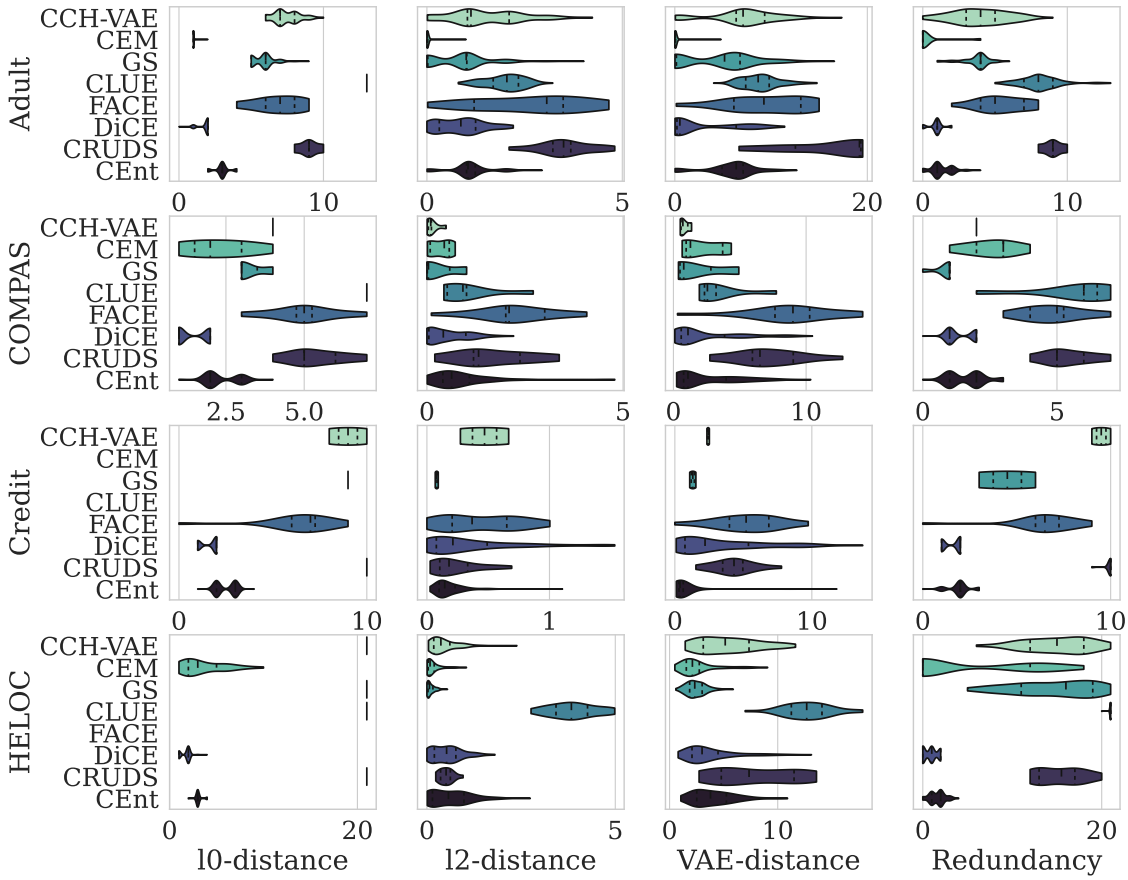


Figure 4.3: Distribution of CEnt's proximity scores across different contrastive methods on the four numerical datasets for LR

tioning that significantly low distance scores, such as in CEM, are an indication of an underlying failure in contrasting the prediction. In our benchmarks, such results were coupled with success rates that can go below 40% for CEM, CCH-VAE, and GS. In such cases, the derived CEs are very close to their original counterparts so they do not flip the decision. CEnt, on the other hand, keeps searching for a counterfactual with no threshold on the distance. This led to consistently low distance measures while maintaining high flip rates. In this sense, it is not surprising that the significantly low redundancy scores attained by CEnt demonstrate the sufficient aspect of the counterfactuals in altering the model's prediction whereas methods such as CRUDS, GS, and CEM can achieve a redundancy score of up to 20 on the HELOC dataset. This implies that most of their derived feature was unnecessary in

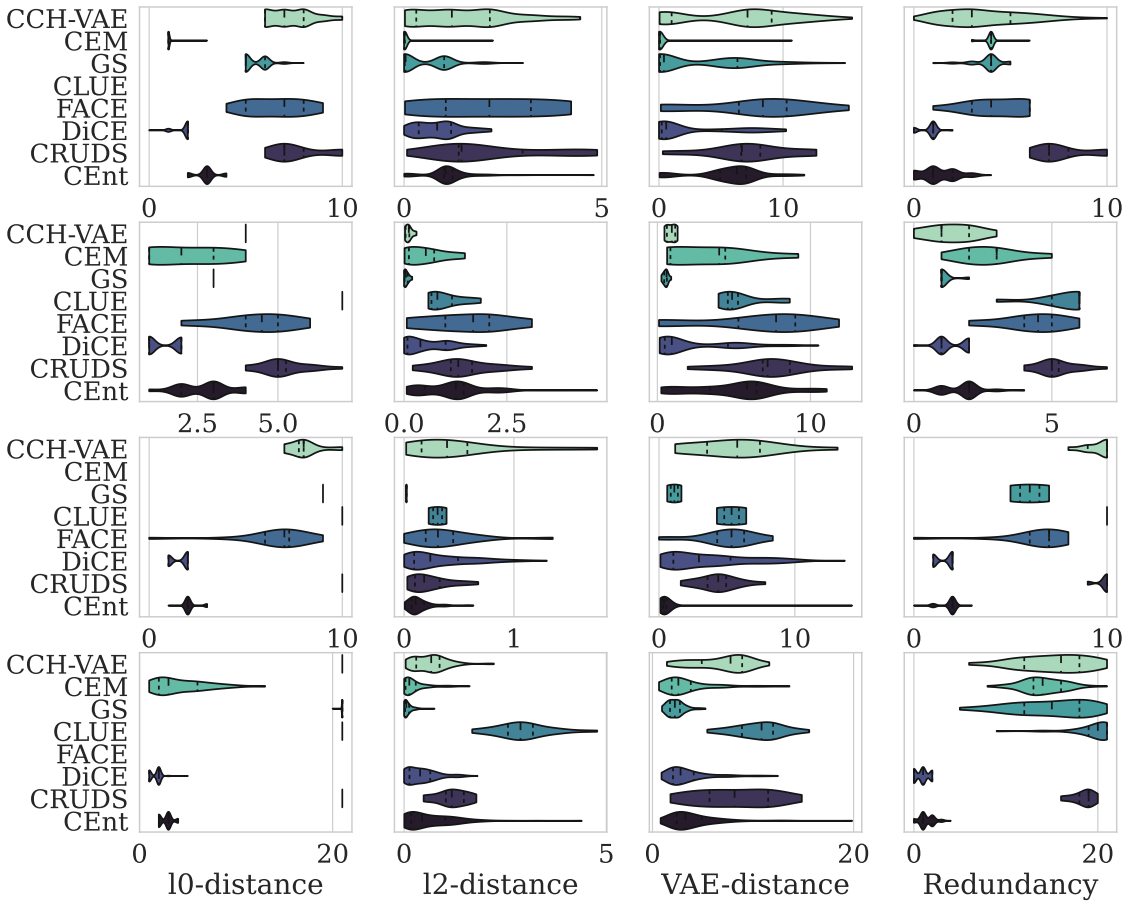


Figure 4.4: Distribution of CEnt's proximity scores across different contrastive methods on the four numerical datasets for ANN

the CE aspect.

4.3.3 *Derivation of Visual Contrasts*

We consider the handwritten digit recognition, MNIST, dataset [144] consisting of 60,000 28x28 images. We consider pixel intensity to be a feature and we derive CEs for the binary classification of confusing digit pairs, i.e. 5 vs. 6, 3 vs. 8, and 1 vs. 9. To this end, we train a CNN with a convolutional layer with 28 units and a 3x3 kernel followed by max pooling and a dense layer of 128 neurons and *relu* activation. A dropout of 0.2 is applied and the activation at the output is *softmax*. We define the visual contrast to be a Gaussian kernel around a pixel whose intensity changed in x'. If the intensity is amplified in x', the contrast is pertinently negative (PN),

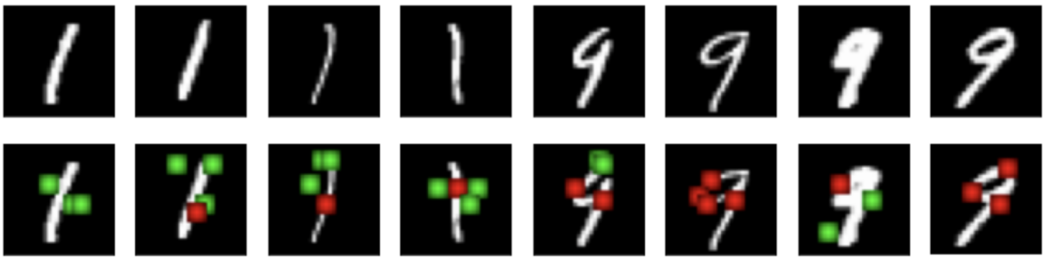
whereas it is pertinent positive (PP) if the intensity is reduced in x'. The visual contrasts in 8 random images for each pair are visualized in Figure 4.5. Generally, for the $5 \rightarrow 6$ contrast, the PNs are the pixels that close the left corner in the lower curve of number 5 and those that make its upper part more curved. PPs are mostly concerned with the upper part of 5. PPs and PNs are reciprocated with the $6 \rightarrow 5$ contrast. More importantly, the derived contrasts are mostly sufficient; i.e. no redundant pixels have been derived an exception in the third image where an outlier region is highlighted in the upper left corner. Interestingly, $3 \rightarrow 8$ is mostly related to the lower curves and not the upper one. This can be explained by the inconsistency in closing the upper loop in 8 (fifth example in Figure 4.5b) whereas the lower one is almost always closed. Similarly, CNN mostly attends to the curve of 9 in the contrast $1 \rightarrow 9$. It is worth mentioning that, in some rare cases, CEnt fails to find a CE; i.e. it derives a CE that does not flip the model's decision. However, in such cases, the contrastive path was, intriguingly, a reasonable and visually appealing contrast even if it did not suffice to change the model's decision.

We select CEM and LIME for a qualitative comparison. Comparison with other methods was not feasible, given their design tailored for tabular datasets. General methods such as FACE and GS are also not compatible with CEnt as the former reports a series of successive examples to reach a CE and the latter's results were not reproducible. Predominantly, explanations derived by LIME, in Figure 4.6, were not useful on MNIST. This can be attributed to the non-contrastive aspect of LIME and its dynamics of operating on super-pixels. The latter reason is crucial; LIME relies on segmentation techniques and is not designed to derive visual contrasts. Additionally, the first three CEs derived by CEM are visually appealing and the last one does not succeed in changing the prediction. However, the visual contrast is not obvious where CEM has the tendency to reshape the digit while flipping a great deal of pixels. CEnt is a minimally invasive process that highlights sufficient

visual contrasts without a major change in shape.

(a) 5 vs. 6

(b) 3 vs. 8



(c) 1 vs. 9

Figure 4.5: Visual contrast with CEnt on MNIST (red represents PPs and green represents PNs)

We also showcase CEnt on the fashion MNIST dataset consisting of 10 different cloth labels for 70,000 28x28 images. We choose 3 pairs of classes that are likely to be mistaken: dress vs. shirt, sandal vs. ankle boot, and T-shirt vs. pullover. We train the same CNN as in the MNIST case and we derive visual contrasts on 8 randomly chosen images in each category in Figure 4.7. The dress vs. shirt contrast is mostly concerned with the sleeves as well as the width of the object. Since dresses are taller than shirts, fitting both in a 28x28 image makes the shirts wider. Intriguingly, CEnt detects this contrast. For sandal vs. ankle boot, CEnt highlights the open holes as

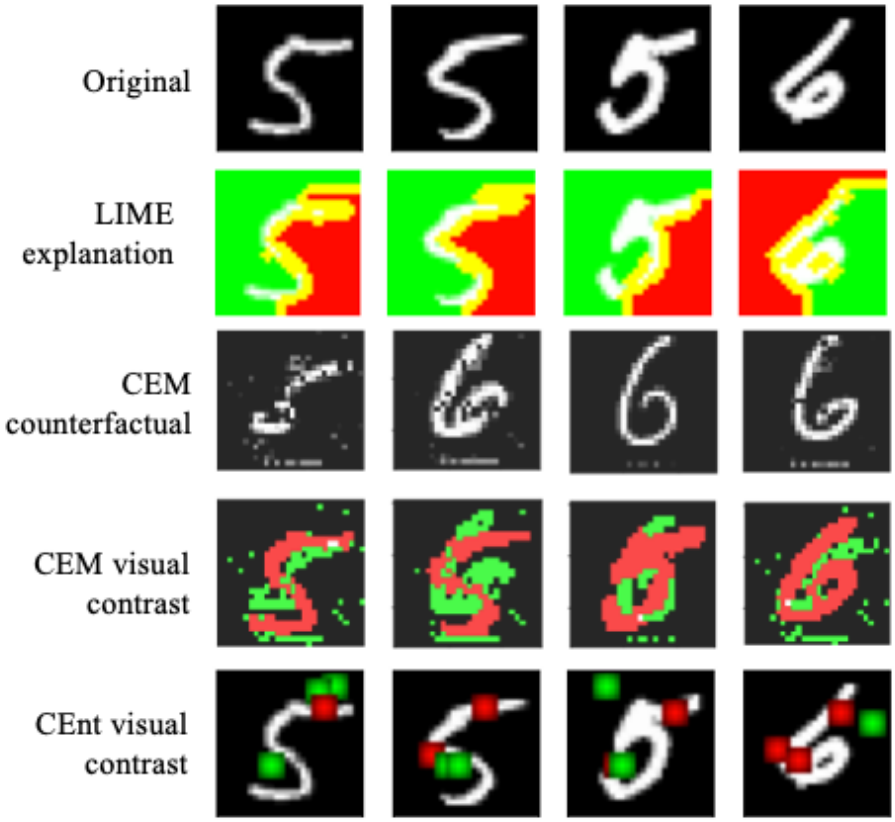


Figure 4.6: MNIST explanations (LIME highlights pixel relevance where red is positive and green is negative, CEM and CEnt highlight contrasts where red is PP and green is PN)

PNs in sandals and the compact areas as PPs in the boots. Finally, CEnt detects the absence/presence of sleeves when contrasting the T-shirt with the pullover.

4.3.4 *Textual Vulnerabilities Detection*

We study an interesting use-case of CEnt in detecting non-useful CEs that serve as adversarial attacks. Instead of employing a VAE distance, we consider a bag-of-words (BoW) approach where sentences are deemed close based on their words with no context, syntax, or semantic integration. Four classifiers were trained on the 20 newsgroup dataset: a random forest, logistic regression, SVM, and a neural network with two fully-connected layers consisting of 100 and 50 neurons. The approximation

(a) Dress vs. shirt

(b) Sandal vs. ankle boot

(c) T-shirt vs. pullover

Figure 4.7: Visual contrast with CEnt on Fashion MNIST (red represents PPs and green represents PNs)

g achieved an accuracy of 98, 93, 96, and 99% respectively. Remarkably, the length of the contrast is on average 1 in all models implying that an insertion or deletion of exactly one word would change the model's prediction. While shorter lengths are an indication of more concise explanations, they do not necessarily reflect the same for textual data when BoW representations. For instance, as displayed in Figure 4.8, including the word "monthly" would change the prediction from *atheism* to *christian* showing the sensitivity to particular words. In this case, CEnt serves as a debugging tool that highlights vulnerabilities in the context of adversarial attacks.

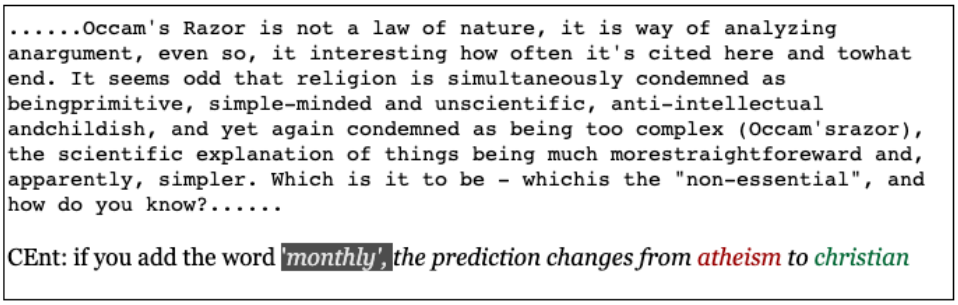


Figure 4.8: CEnt on an instance of the 20 newsgroup data

4.4 Research Directions

In this objective, we develop a wide plan of attack for algorithmic recourse by accounting for custom costs and elegantly addressing semi-immutability and plausibility. We propose CEnt, a novel entropy-based method that supports an individual facing an undesirable outcome under a decision-making system with a set of actionable alternatives to improve their outcome. CEnt samples from the latent space learned by VAEs and builds a decision tree augmented with feasibility constraints. Graph search techniques are then employed to find a compact set of feasible feature tweaks that can alter the model's decision.

Our empirical evaluation on real-life datasets shows improvement in proximity, latency, and attainability without constraint violation. CEnt has immense potential in adapting to non-tabular data where it identifies visual contrasts and serves as a debugging tool to detect the model's vulnerabilities.

CEnt's results suggest several future directions. First, the exploration of different data representations, such as embeddings or super-pixels, improves robustness and widens the applicability of CEnt on different data types. Second, we wish to improve the privacy guarantees of our method by exploring generative data techniques to alleviate the need to access training data.

Furthermore, the derived contrastive examples can be used to detect discrimination cases in classifiers. In the next objective, we define contrastive fairness, and we

augment it with faithfulness guarantees. The latter is inspired by the attainability of the contrastive sets used in the comparison during the fairness evaluation.

Chapter 5

Objective II: Faithful Contrastive Fairness

In an attempt to detect discrimination in datasets and classifiers, researchers first tried to define *fairness*. The result was more than twenty different definitions as surveyed in Section 3.2.1. In what follows, we outline discrimination cases that existing fairness metrics cannot directly detect. Those cases are (1) non-tabular datasets settings were protected (sensitive) attributes are not explicitly reported, (2) situations where sensitive attributes are a *legacy*, (3) the comparison of the individual to an unattainable counterpart, and (4) the reliance on distance metrics that do not reflect real-world similarity-based measures. Accordingly, we present Faithful Contrastive Fairness (FCF) as an individual fairness metric that considers concentric neighbors around an individual and computes the corresponding contrastive cost. A prediction is considered fair if these costs do not entail a change in the protected attribute or its dependents. More importantly, the neighborhood is derived based on a manifold-like distance metric computed by auto-encoders that account for data density and attainability.

We look at the *faithfulness* of contrastive explainability that defines three conditions to design useful explanations, and these are proximity, connectedness, and

stability [39], [40]. We motivate these properties in fairness settings and apply them to contrastive fairness.

This chapter presents the second objective of our dissertation and is organized as follows. We introduce the problem statement in Section 5.1. Then, we propose the methodology behind FCF, our faithful *contrastive* fairness, and its computation in Section 5.2. We discuss *faithfulness* guarantees in Section 5.3 before validating on synthetic and real-world datasets in 5.4 and concluding with future directions in Section 5.5.

5.1 Problem Statement

We consider the following cases not covered in existing fairness metrics.

Case 1: *The adjustment of protected attributes is impractical. This can be seen in text and images or in tabular datasets where the protected attribute is removed from the data without its dependents. As an example, one can infer race from certain neighborhood. It is not clear how to perturb the race feature when it is removed from the training data, while its dependent (neighborhood) is not.*

Case 2: *The protected attribute is a legacy. Solely, it cannot change the model's decision; but when coupled with other attributes the decision might change.*

Case 3: *Altering protected attributes leads to unattainable cases when the occurrence probability is not considered. In this case it is unfair to detect discrimination by comparing an individual to a better-treated unattainable counterpart.*

Case 4: *Devising meaningful distance measures is difficult or it does not reflect real data similarities.*

Motivated by these scenarios, we investigate a novel definition of fairness that can account for the aforementioned cases. We denote the D-dimensional feature space as $\mathcal{X} \subseteq \mathbb{R}^D$. For an observation, i, is represented by a feature vector $x_i \in \mathbb{R}^d$. The protected attribute a_p and its dependents a belong to a set $DEP(a_p)$. We

assume a pre-trained classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ such that $\mathcal{Y} = \{c_1, c_2, \ldots, c_k\}$. *$f$ can be a binary classifier with* $\mathcal{Y} = \{0, 1\}$. We study the fairness of f with respect to a set of protected attributes A without making any assumptions about f.

5.2 Contrastive Fairness

We represent our notion of fairness by FCF: **F**aithful **C**ontrastive **F**airness and we explain its faithfulness later in Section 5.3. Our definition starts from a particular observation, considers concentric *spheres* around it with a predefined radius of ϵ, and searches for contrastive examples within these spheres. The cost of these proximate contrasts is then checked against the dependency on the protected attributes.

For $\epsilon \in \mathbb{R}$, we define the $\epsilon-$*contrastive neighborhood* of observation i represented by x_i as:

$$\mathcal{N}_\epsilon^{CT}(x_i) := \{x_j \in \mathcal{X} | \quad d(x_j, x_i) \leq \epsilon \text{ and } f(x_j) \neq f(x_i)\} \tag{5.1}$$

Accordingly, the contrastive cost can be defined as the change in the input features between the input and instances in the contrastive neighborhood. Formally, the contrastive cost of an individual x_i is $\overrightarrow{x_i x_j} \quad \forall x_j \in \mathcal{N}_\epsilon^{CT}(x_i)$. Having defined the main components, it remains to define our contrastive fairness metric as follows.

Definition 7 (Faithful Contrastive Fairness)**.** *A predictor f is fair if the cost of any contrastive neighbor (derived faithfully) does not encode the protected attribute and its dependents. Formally, f is fair if*

$$\overrightarrow{x_i x_j} \perp \overrightarrow{a} \quad \forall a \in DEP(a_p)\ \forall x_j \in \mathcal{N}_\epsilon^{CT}(x_i) \tag{5.2}$$

In the case of tabular data, $\overrightarrow{a} = e_q$ where $e_q[l] = 0 \forall l \neq q$ and $e_q[q] = 1$; q is the index of a in the list of attributes. If a is not implicit in the dataset, i.e. $\overrightarrow{a} = e_q$ cannot be easily established, we generalize the definition above to:

Definition 8. *(Generalized Faithful Contrastive Fairness) A predictor f is fair if $\forall a \in DEP(a_p)$ and $\forall x_j \in \mathcal{N}_\epsilon^{CT}(x_i)$, a cannot be inferred from the contrastive cost $\overrightarrow{x_i x_j}$.*

Inference can be tested by looking at causal graphs or by training a separate classifier $\hat{g}(\overrightarrow{x_i x_j}) = a$ in the case of images and text.

5.2.1 *Latent distance*

Instead of measuring $d(x_j, x_i)$ as the (normalized) l_1 or l_2 norms for numerical features or a simple matching distance for categorical features; we exploit the structure that latent-variable models learn. We base our observation on the assumption that high-dimensional instances can be better represented by points in a much simpler latent space. Simpler measures, such as the Euclidean distance of the cosine similarity can be used in the latent space [40] as it is considered to be Euclidean [103], [145]. Hence, we consider the latent representation $z_j \in \mathcal{Z} \subset \mathbb{R}^{d_2}, d_2 < d$ and we measure $d(x_j, x_i)$ as the latent distance or L-p norm between z_j and z_i.

Latent representation distance is thus a manifold-like distance that reflects (1) data distribution, (2) attainability, and (3) application-specific distance notion that is reflected from the input data. We compute latent representation through Variational Auto-Encoders (VAEs) as detailed in Section 4.2.1. More precisely, given an observation $\boldsymbol{x}$, its latent representation $\boldsymbol{z}$ can be learned in a self-supervised manner.

To this end, we define an encoder $m(.)$ with its parameters Φ and distribution $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$ and a decoder $g(.)$ parametrized by Θ with a likelihood of $p_\Theta(\boldsymbol{x}|\boldsymbol{z})$. Φ and Θ are represented by non-linear functions. We then define $p_\mathcal{D}(\boldsymbol{x})$ as the empirical distribution of the data. Under these circumstances, the evidence lower bound (ELBO) [103] can be used to compute the intractable integral above as:

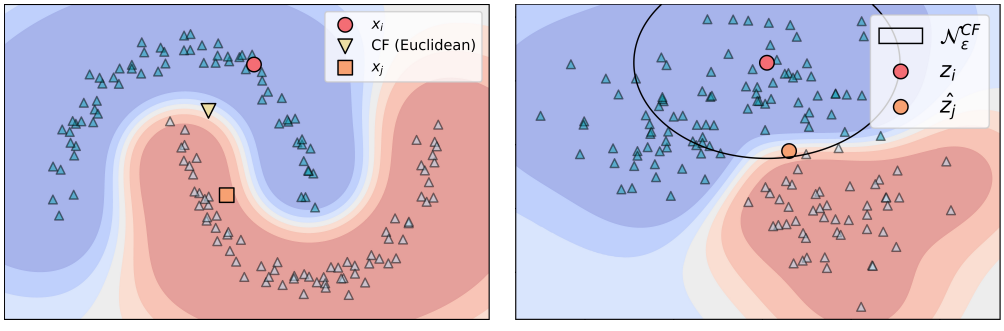
$$\mathbb{E}_{p_\mathcal{D}}(\boldsymbol{x})\Big[\log p_\Theta(\boldsymbol{x})\Big] \geq \mathbb{E}_{p_\mathcal{D}}(\boldsymbol{x})\Big[\mathbb{E}_{q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\Theta(\boldsymbol{x}|\boldsymbol{z})\big] - \mathbb{KL}\Big(q_\Phi(\boldsymbol{z}|\boldsymbol{x})|p(\boldsymbol{z})\Big)\Big] \quad (5.3)$$

where $\mathbb{KL}$ is the relative entropy or the Kullback–Leibler divergence [138]. Moreover, $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\Theta(\boldsymbol{x}|\boldsymbol{z})$ are assumed to be Gaussian. In the case of binary attributes, the decoder can be assumed to be Bernoulli. The latent representation $\boldsymbol{z}$ can be found by optimizing the right-hand side of the equation above.

5.2.2 *Computation of FCF*

We assume a pre-trained, potentially non-linear, predictor f parametrized by $\boldsymbol{\theta}$ such as a neural network. Given a dataset $\mathcal{D} \equiv \{(x_l, y_l)\}$ for $l = 1, \ldots, n$, the decoder parameters Φ can be learned as described in the previous section.

Algorithm 1 describes the dynamics of FCF computation. First, the VAE is trained and the latent representation z_i is accordingly derived. Then, concentric neighborhoods centered around z_i in $\mathcal{Z}$ are considered. For each neighborhood, J points are sampled and filtered according to f's prediction; only those that change f's decision are kept. Then, the counterfactual cost is checked for dependence on the protected attribute. The algorithm returns the average of the bias degree in the neighborhood. The illustration is shown in Fig. 5.1 which also contrasts the Euclidean neighborhood to $\mathcal{N}_\epsilon^{CF}$.



(a) Input space $\mathcal{X}$ (b) Latent space $\mathcal{Z}$

Figure 5.1: Illustration of Algorithm 1. The neighborhood is computed in the latent space $\mathcal{Z}$ and mapped back to the original space $\mathcal{X}$. For illustration purposes, $d = d_2 = 2$.

Algorithm 1 Counterfactual Fairness Evaluation

Input: classifier f_θ, observation data $\mathcal{D}$, neighborhood radius ϵ, increments δr, protected attribute a_p and i^{th}observation x_i

Initialize:

$\Phi \leftarrow \arg\max \mathbb{E}_{p_\mathcal{D}}(\boldsymbol{x})\left[\mathbb{E}_{q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\Theta(\boldsymbol{x}|\boldsymbol{z})\right] - \mathbb{KL}\left(q_\Phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right)\right]$

$z_i \leftarrow m(x_i)$

$r \leftarrow 0$

repeat

  $r = r + \delta r$

  Initialize $\mathcal{N}_r^{CF}$ in latent space

  for $j = 1$ **to** J **do**

    Sample $\tilde{z_j}$ based on perturbation on z_i with δr

    $\tilde{x_j} = g(\tilde{z_j})$

    if $f(\tilde{x_j}) \neq f(x_i)$ **then**

      $c = c + 1$

      $s = s + |\cos(\overrightarrow{x_i\tilde{x_j}}, \overrightarrow{a_p})|$

    end if

  end for

  $\text{bias}_r = s/c$

until $r \geq \epsilon$

return $\text{bias}_\epsilon = \overline{\text{bias}_r|_{r=0}^{\epsilon}}$

Considerations Algorithm 1 checks for orthonormality between $\overrightarrow{x_i\tilde{x_j}}$ and $\overrightarrow{a_p}$ by computing the cosine of the angle between them. When the two vectors are orthonormal their $\cos(.,.) = 0$, and the bias will not be affected. Otherwise, we consider the $\cos(.,.)$ as a continuous measure of non-orthonormality. We note, that, in the case of categorical attributes $\cos(\overrightarrow{x_i\tilde{x_j}}, \overrightarrow{a_p})$ will either be 0 or 1. Furthermore, the algorithm assumes explicit protected attribute encoding a_p. If not, the statement $s = s + \cos(\overrightarrow{x_i\tilde{x_j}}, \overrightarrow{a_p})$ can be replaced with an estimate of $\cos(.)$ based on the probability of inferring a_p from $\overrightarrow{x_i\tilde{x_j}}$. Finally, the algorithm returns $\overline{\text{bias}_r}$, the average of the bias found in $\mathcal{N}_\epsilon^{CF}$. This average can be weighted by the inverse of $\mathcal{N}_\epsilon^{CF}$ radius, i.e. the series $1/\delta r, 1/(2\delta r), \ldots, 1/\epsilon$. This series reflects similarities in $\mathcal{Z}$ where closer counterfactuals have a higher impact on FCF.

From continuous to discrete measure of fairness Algorithm 1 computes

the (weighted) ratio of the contrastive neighbors that have a different value for the protected attribute. Therefore, one can infer that a model is biased if $\overline{\text{bias}_\epsilon} > 0$. Nevertheless, we can relax the above condition to $\overline{\text{bias}_\epsilon} > t$.

5.2.3 *Relation to Existing Notions*

Relation to counterfactual fairness of [31]

FCF is not a causal metric; it is instead an individual fairness evaluator that compares two proximate individuals while searching for an underlying bias. The implementation of counterfactual fairness in [31] trains a fair classifier on mutated inputs. The mutation is applied through a de-convolution process from a causal perspective. Defining $\mathcal{N}_\epsilon^{CF}(x_i)$ as:

$$\mathcal{N}^{CF}(x_i) := \{x_j \in \mathcal{X} \quad | \quad x_j = x_{i,A \leftarrow a'} \text{ and } f(x_j) \neq f(x_i) \forall a' \text{ attainable by } A\} \tag{5.4}$$

would make the definition of [31] and FCF similar. However, we do not restrict the counterfactuals to perturbations in a_p only, which allows FCF to cover cases 2 and 4 mentioned in Section 5.1. Moreover, we generate counterfactuals then we infer dependency which solves the issue discussed in case 1. Finally, our computation of the latent distance considers the occurrence probability and the attainability, targeting case three.

We draw the reader's attention that contrastive sets can be generated for compound datasets such as images as in [75], [146], text as in [147] and graphs as in [148]. The existence of such methods makes FCF applicable to a wide range of data types.

Relation to fairness through unawareness [149]

Fairness through unawareness considers a model fair if its outcome does not change

had the protected attribute been hidden from the model. FCF extends this definition to cases where such attributes are not explicit. This is crucial, especially since discrimination cases can occur even when the model is unaware of the sensitive data. FCF can cover those cases and can further measure the degree of the model's *awareness* of some sensitive attributes.

Relation to individual fairness [38]

FCF is closely connected to individual fairness, which requires that if two individuals are *close* in the feature space, their prediction should also be *close*. FCF is a *de-compiled* individual fairness. We take the neighborhood of a point, and we filter instances that have a different prediction. We study those neighbors with the lens of dependence on protected attributes. On the other hand, individual fairness is agnostic with respect to the notion of similarity metric, which is a double-edged sword: it can generalize well but it can have unfavorable outcomes when there is no unified way of defining similarity. Our contrastive fairness, FCF, measures similarity based on latent representations which account for data density distribution, occurrence, and attainability.

5.2.4 *Extension to Group Fairness*

Group fairness requires that two groups (privileged and underprivileged) have the same probability of being assigned a favorable outcome. FCF can be extended to group fairness by computing the expected value of the bias between different sampled individuals.

Definition 9. *[Demographic Faithful Contrastive Fairness] A predictor is fair to an underprivileged group* $g_{underprivileged}$ *if*

$$\mathbb{E}_{g_{underprivileged}}\Big[bias_{\epsilon}\Big] = \mathbb{E}_{g_{privileged}}\Big[bias_{\epsilon}\Big] \tag{5.5}$$

We relax the criterion by considering the predictor to be fair if

$$\sigma < \frac{\mathbb{E}_{g_{\text{underprivileged}}}\left[\text{bias}_\epsilon\right]}{\mathbb{E}_{g_{\text{privileged}}}\left[\text{bias}_\epsilon\right]} < \frac{1}{\sigma} \tag{5.6}$$

for $0 < \sigma \leq 1$. σ values that are closer to 1 entail stricter fairness requirements. Additionally, for $\sigma = 0.8$, our definition of group fairness would satisfy the four-fifths rule. This rule states that if the selection rate for a particular group (e.g underprivileged) is less than 80% of that of the group with the highest selection rate (e.g. privileged), there is an adverse impact on that group [150].

5.3 Faithfulness Guarantees

The most intuitive desideratum for individual fairness is to be *faithful* to the examined individual. This entails a *"fair"* comparison to individuals that are **close** and **attainable**. Consequently, the generated $\epsilon-$contrastive neighbors are required to be generated in a (1) stable manner while satisfying (2) proximity and (3) connectedness. In this section, we present the first extension of the *contrastive faithfulness* notions of explainability [39], [40], [151] to individual fairness and we present quantification metrics.

Next, we assume $0 \leq b(x_i) \leq 1$ to be a continuous measure of the bias degree for observation x_i and $1 - b(.)$ to be its corresponding fairness metric.

5.3.1 *Stability*

The fairness measure should be coherent locally by forcing a close $b(.)$ for close neighbors. Formally,

Definition 10 (Stability of fairness metric)**.** *A fairness metric* $(1 - b(.))$ *is stable if*

$$\frac{d_1(b(x_{i_1}), b(x_{i_2}))}{d_2(x_{i_1}, x_{i_2})} < M \tag{5.7}$$

We suggest $d_1(c_1, c_2) \equiv |c_1 - c_2|$ and $d_2(.,.)$ derived by VAEs as in Section 5.2.1 and propose the following measure of stability.

$$\frac{d_1(b(x_{i_1}), b(x_{i_2}))}{d_2(x_{i_1}, x_{i_2})}. \tag{5.8}$$

5.3.2 *Proximity*

Proximity ensures that the contrastive neighbor is not an exception, i.e. it is attainable. Hence, proximity measures whether the considered neighborhood is outlying with regard to ground truth data.

Definition 11 (Proximity in contrastive fairness)**.** *A contrastive fairness metric satisfies proximity if the distance between the considered neighbor x_j and x_i is proximate to the distance between x_i and* $\underset{x_k \in \mathcal{X}, f(x_k) \neq f(x_i)}{\arg\min} d(x_i, x_k)$.

Proximity can be measured by

$$\frac{d(x_i, x_j)}{\min\limits_{x_k \in \mathcal{X}, f(x_k) \neq f(x_i)} d(x_i, x_k)} \tag{5.9}$$

which is reflected by the Local Outlier Factor (LOF) as in [152].

5.3.3 *Connectedness*

Connectedness ensures that the derived contrastive neighborhood is continuously connected to a ground-truth observation of the same class using the topological notion of the path. We borrow the below definition from [39].

Definition 12 ($\epsilon-$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 $\forall i < Nd(e_i, e_{i+1}) < \epsilon \; \forall n < N, f(e_i) = f(e)$.

Connectedness is a binary indicator, i.e. it is 1 if x_j is ϵ-connected to x_i and 0 otherwise. Computing connectedness is complex. However, two points can be considered connected if they belong to the same cluster computed by DBSCAN algorithm [153] with the min points parameter set to 2 [39].

In NLP settings, we use latent space embeddings encoded by language models such as GPT-2 [42] and we compute their cosine similarity. We leave faithfulness in imagery settings as a future extension of FCF.

5.4 Results

In this section, we showcase the motivation behind FCF on a synthetic dataset. Then, we study the applicability and the *faithfulness* of FCF on real-world datasets and we compare it to existing individual and group fairness metrics.

The experiments are run on an Intel(R) Core(TM) i7 machine with a 4-core CPU with *Python 3.7.4* and *scikit* backend. Each experiment is repeated 5 times and the average performance is reported.

5.4.1 *Synthetic Experiment Setup*

This experiment highlights the fundamentals behind FCF. We demonstrate how the comparison to proximate attainable neighbors by FCF serves fairness purposes.

We demonstrate how our counterfactual fairness metric addresses cases 2 and 3 in a synthetic experiment. We generate 1200 points equally distributed between the privileged and underprivileged groups. We consider two continuous features x_1 and x_2 following quadratic (noisy) functions. We consider a simple classifier $f, f(x_1 = ., x_2 > 0) = 1$, and 0 otherwise. We also assume a protected (continuous) attribute x_2 that is not causally related to x_1. Fig. 5.2 shows the dataset and the classification boundary. More details on data generation are provided in Appendix 8.

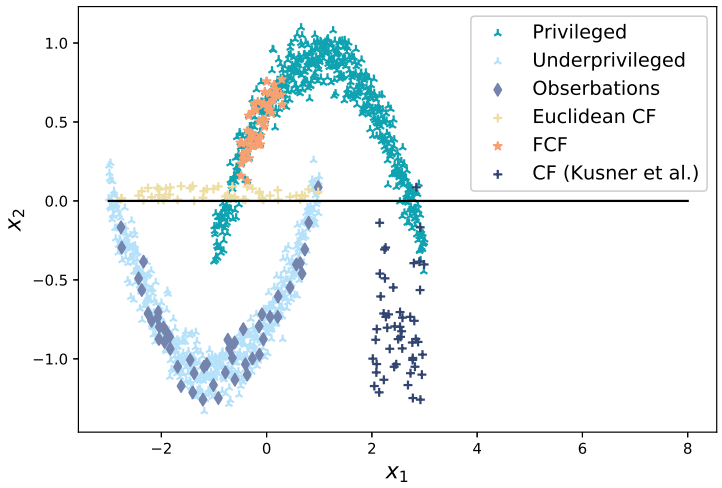


Figure 5.2: Synthetic experiment with original data distribution and the generated counterfactuals

We sample 50 observations and compute their contrastive neighborhood (or counterfactuals) based on CF where only the sensitive attribute ($x1$) is altered. Additionally, we consider the counterfactuals derived by the Euclidean-like recourse methods (Euclidean CF) such as actionable recourse [24] and contrast them with FCF's counterfactuals.

We illustrate the neighborhood in the original space $\mathcal{X}$ and the latent space $\mathcal{Z}$ in Fig. 5.3 where the latter space is used by FCF.

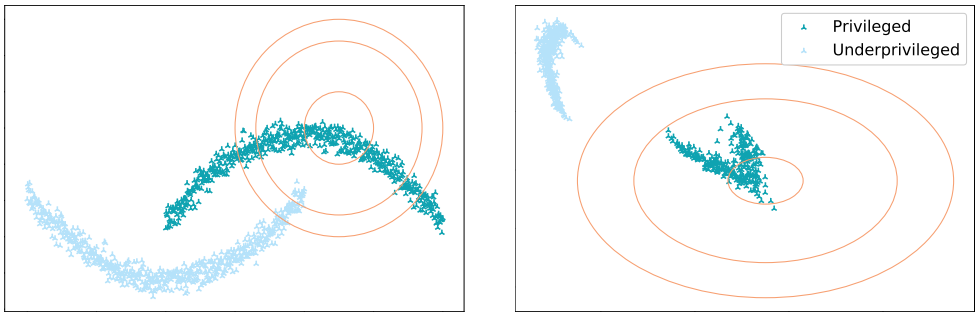
5.4.2 *Synthetic Experiment Results*

As shown in Fig. 5.2, the contrastive neighbors generated by CF are not attainable based on the data distribution. Second, the decision for the counterfactuals of CF remains unchanged (all points are still below the classification boundary), implying that f is fair. Euclidean CF advises a change in x_2 (not sensitive) that leads to unattainable points. One cannot consider that f is fair if the evaluation depends on unattainable counterfactuals. On the contrary, FCF that considers VAE distances leads to attainable contrastive neighbors that belong to a different demographic group. Those neighbors are obtained by an alteration on x_1 and x_2 combined. The

computation of FCF as explained in Algorithm 1 yields scores < 1 hinting to bias cases as shown in Fig. 5.4.

However, CF leads to scores that are mostly 1 (Fig. 5.4) missing thus obvious discrimination cases. On the other hand, while this bias is detected by Euclidean CF, the detection cannot be reliable as it is inferred based on unattainable contrasts.

That being the case, FCF can help marginalized groups *faithfully* prove discrimination cases that can be hard to unveil with other metrics.



(a) $\mathcal{X}$ space (b) $\mathcal{Z}$ space

Figure 5.3: Latent space visualization of the neighborhood

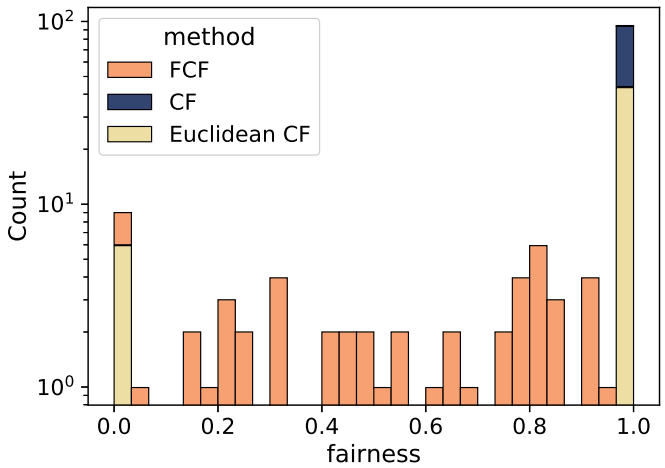


Figure 5.4: Fairness distribution following a log scale for the y-axis

More on attainability

We further highlight the attainability of contrastive neighborhoods by running state-

of-the-art methods on the German credit scoring dataset[1] and training a Logistic Regression (LR) and an Artificial Neural Network (ANN) model. Experimental details are provided in Appendix 8. The task is to classify loan applicants as having good or poor credit risks. The bias is checked against age and gender. To highlight attainability, we only consider age in this experiment. Counterfactual explainability methods are CEM [154], CLUE[155], DICE[89], GS [156] and Wachter [23].The results in Table 5.1 show that, with Euclidean-like distances, contrastive neighbors might be unattainable as the average age cost is already beyond the life span and this does not comply with "commonsense constraints". FCF, on the other hand, derives a contrastive neighborhood with attainable age costs that respect the data distribution. Comparison based on an attainable and plausible contrastive neighborhood can thus be deemed faithful. Although some methods can take the constraint into consideration by setting the attribute to immutable, we disabled this option as FCF necessitates an alteration of protected attributes.

	CEM	**CLUE**	**DICE**	**GS**	**Wachter**	**FCF**
LR	94	1796	93	94	1798	29
ANN	2307	86	2305	2307	2291	32

Table 5.1: Counterfactual age cost for the German credit scoring dataset

5.4.3 *Real World Experiment Setup*

In this set of experiments, we show how FCF detects discrimination cases missed by other metrics, and we demonstrate the faithfulness of our approach.

We test our metric on three popular real-world datasets in the fairness literature: the aforementioned German credit scoring, the adult census income dataset [2] and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset. The adult census dataset is used to predict whether an adult's

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[2] https://archive.ics.uci.edu/ml/datasets/adult

income $> 50K$ USD where race and gender are sensitive attributes. The goal in COMPAS is to score the defendant's likelihood of reoffending where race is the protected attribute. We consider categorical sensitive attributes (gender and race) and continuous attributes (age) in the credit data.

We employ 60% of each dataset to train the model, 20% to validate the model and select the hyperparameters, and 20% to test the FCF scores. We experiment with a Decision Tree (DT), Support Vector Machines (SVM) model, Logistic Regressor (LR), and Artificial Neural Network (ANN). DT is trained on Gini impurity with a maximum depth of 5 and SVM is trained with an RBF kernel. LR is a $l2$-penalized model and we consider $d_2 = 5$. ANN consists of 2 hidden layers of 13 and 4 neurons activated via ReLU and trained using a weighted binary cross-entropy loss function through gradient-descent with root mean squared propagation for 50 epochs with a 10^{-3} learning rate.

The VAE consists of 2 hidden encoder layers with 16 and 8 neurons, respectively, and 10 neurons in the bottleneck layer. We train the VAE for 50 epochs with a batch size of 64 and a learning rate of 10^{-3}. We select the VAE architecture based on the loss of validation data using a grid search.

5.4.4 *Real World Experiment Results*

Detected Bias We sample 100 instances from the validation data set and compute their FCF scores. We consider a model to be biased against an individual if the FCF bias score is > 0.1. The distribution of the scores is shown in Appendix 8.

	Adult		**German**		**COMPAS**
	race	**gender**	**race**	**gender**	**race**
DT	4	7	5	2	3
SVM	5	9	11	7	4
LR	4	5	10	5	2
ANN	6	9	7	7	5

Table 5.2: Number of individuals discriminated against according to FCF

We can see that FCF detects discrimination against race and gender in the three datasets and for the four models. We validate the faithfulness of our process and we compare it to existing work in the next part.

5.4.5 *Comparison to Existing Metrics*

We consider the adult data set to compare the FCF with existing group fairness metrics. Mainly, we compute the equality of opportunity (EqOpp), predictive parity (PredP), predictive equality (PredE), and statistical parity (StatP). Furthermore, we calculate the FCF for privileged and underprivileged groups and report the ratios of their expected values as in Equation 5.6. Since the Adult dataset conveys two protected attributes, race and gender, we consider the privileged group to be `white-male` and the underprivileged group to be `while-female`, `black-female` and `black-male`. We report group fairness results in Table 5.3 and we underline discrimination cases where the score does not fall within the $[\sigma, \frac{1}{\sigma}]$ range with $\sigma = 0.9$. We select σ as such to enforce fairness guarantees stronger than the four-fifths rule.

	Acc	**EqOpp**	**PredP**	**PredE**	**AccEq**	**StatP**	**FCF**
DT	84	1.00	1.02	0.99	0.98	0.99	0.91
SVM	85	1.02	0.98	1.10	1.01	1.00	<u>1.12</u>
LR	83	<u>0.89</u>	<u>1.22</u>	1.02	0.97	1.02	<u>1.23</u>
ANN	86	0.91	<u>1.13</u>	1.02	<u>1.10</u>	1.03	<u>1.20</u>

Table 5.3: Group fairness metrics along with FCF on the Adult dataset where underlined scores indicate discrimination with $\sigma = 0.9$

As shown, FCF detects discrimination with the SVM, LR, and ANN models. These cases were detected mainly by equality of opportunity and predictive parity but were missed by predictive equality and statistical parity. Next, we study whether the bias cases are faithfully derived.

5.4.6 *Faithfulness Results*

Faithfulness is studied in contrastive learning settings. Thus, its evaluation is feasible only for FCF and its CF counterpart [31].

Stability

To check the stability of FCF, we performed perturbations on $\mathcal{Z}$, and noted the change in FCF. Fig. 5.5 shows that the change in FCF is bounded above by a linear function when the alteration $\epsilon \leq 0.5$, implying a stable algorithm. We note that computing the stability of CF is trivial where $d(x_i, x_j)$ in Equation 5.8 is always 1 because CF's counterfactuals are alterations on the protected attribute only.

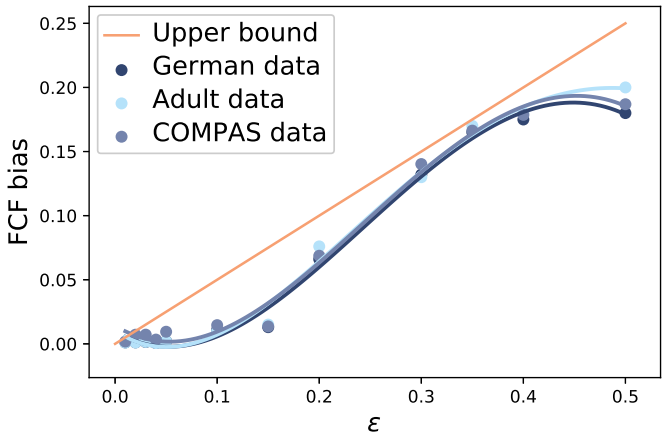
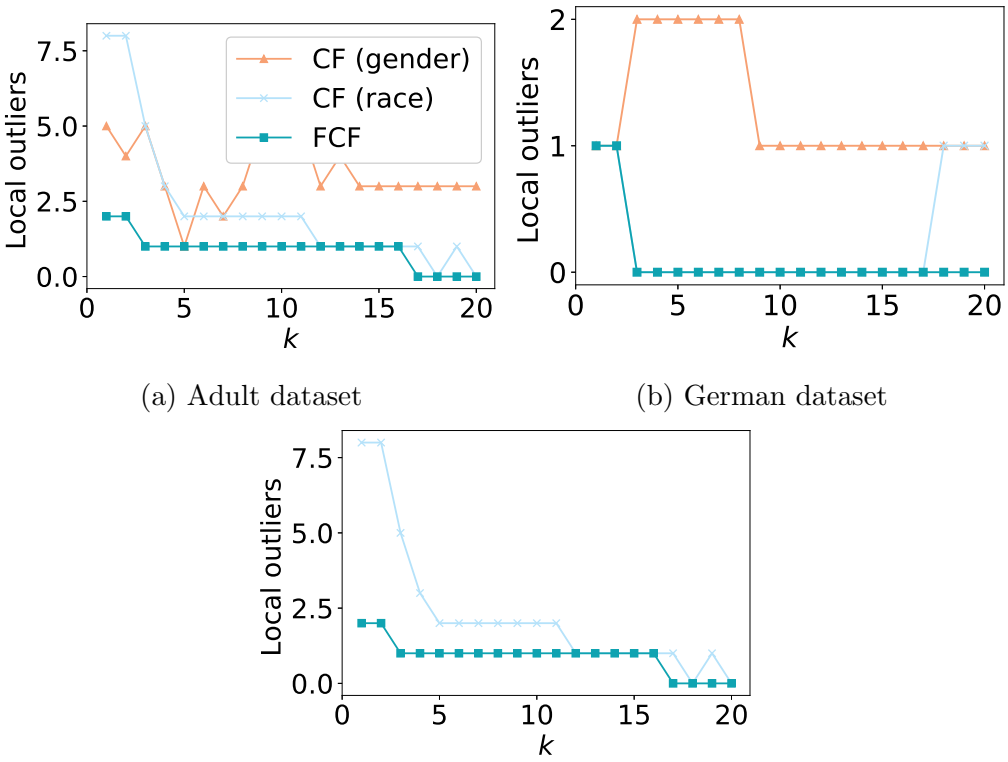


Figure 5.5: Stability score of FCF on the German and adult datasets

Proximity and Connectedness

The scores are shown in Fig 5.6 and Fig 5.7. Connectedness is represented by the percentage of not connected contrastive neighbors while changing ϵ of DBSCAN. The proximity is inversely proportional to LOF by changing the number of neighbors k. Both FCF and CF achieve very high connectedness scores with nearly 0 unconnected neighbors. However, FCF outperforms CF in proximity score where contrastive neighbors generated by FCF are less likely to be outliers.

Thus, FCF is shown to be *faithful* for the individual examined for fairness. The



(a) Adult dataset

(b) German dataset

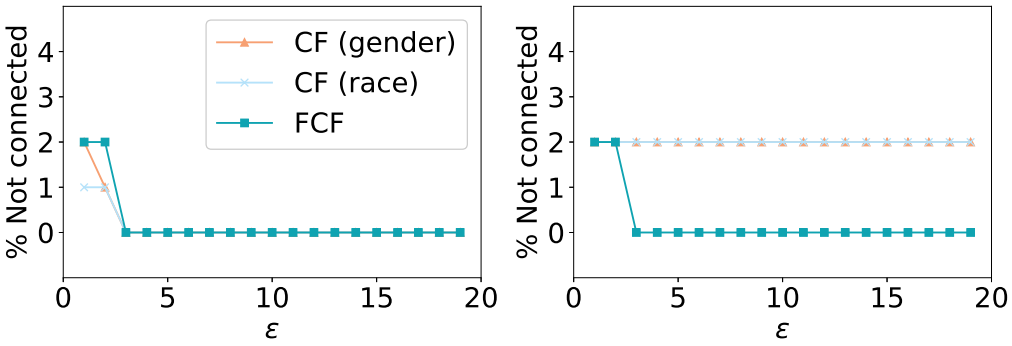
(c) COMPAS dataset

Figure 5.6: Proximity scores (low numbers of local outliers are desirable)

conducted experiments show how the individual is compared, in a stable way, to contrastive neighbors that are proximate and connected.

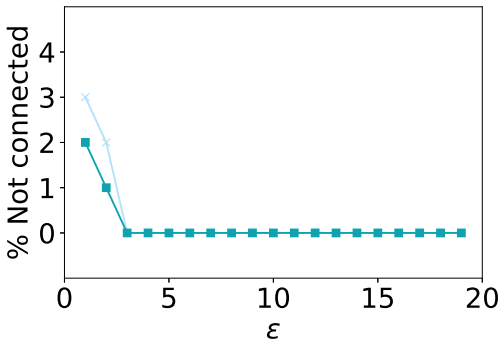
5.4.7 *Impact of VAE Architecture*

Finally, we conduct a qualitative and quantitative study on the impact of the VAE complexity and the dimension of the bottleneck layer on the FCF scores and their faithfulness. Intuitively, we prefer a powerful VAE (deeper layers) that can learn the underlying data distribution. However, a deep one trained on a non-complex dataset with linear relationships might simply learn to copy its inputs to the output, without learning any meaningful representation. Thus, the loss of the VAE on newly seen data should be the main factor in the validation process to select complex, yet generalizable auto-encoders that can reflect the data distribution.



(a) Adult dataset

(b) German dataset



(c) COMPAS dataset

Figure 5.7: Connectedness scores (low %not connected scores are desirable)

Second, we consider the size of the bottleneck layer which implies the dimension of the latent representation. Few nodes in the bottleneck might yield to an under-representative learned manifold; whereas a significantly increased number of nodes can lead to over-fitting.

To showcase this, we train the following VAEs:

- VAE: the default VAE in this work with 2 hidden encoder layers with 16 and 8 neurons and 10 neurons in the bottleneck layer.
- Deep-VAE: a deeper VAE with 10 hidden encoder layers with 64, 50, 40, 32, 16, 10, and 10 neurons in the bottleneck layer.
- Shallow-VAE: a shallower VAE with 1 hidden encoder layer with 16 neurons and 10 neurons in the bottleneck layer.

- VAE-B1: 2 hidden encoder layers with 16 and 8 neurons and 1 neuron in the bottleneck layer.
- VAE-B100: 2 hidden encoder layers with 50 neurons in each and 100 neurons in the bottleneck layer.

We consider the Adult dataset and we compute the group scores of FCF on the latent representations computed by the aforementioned VAEs. We report the FCF scores in Fig. 5.8 where fair scores, that is, between 0.9 and 1/0.9, would belong to the green region.

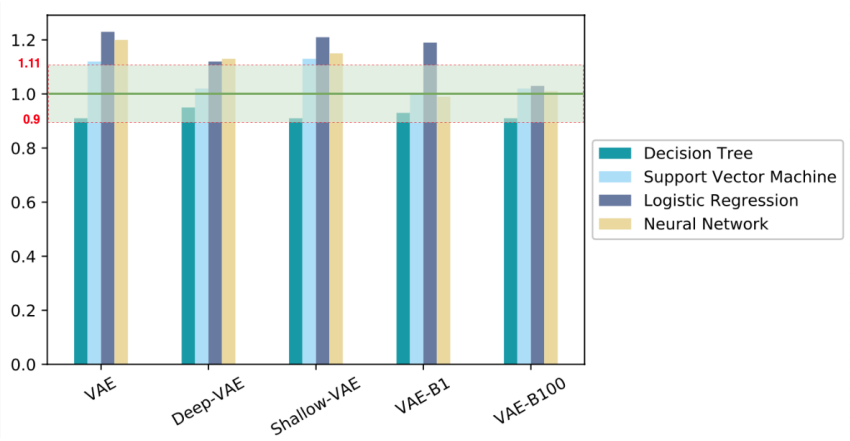
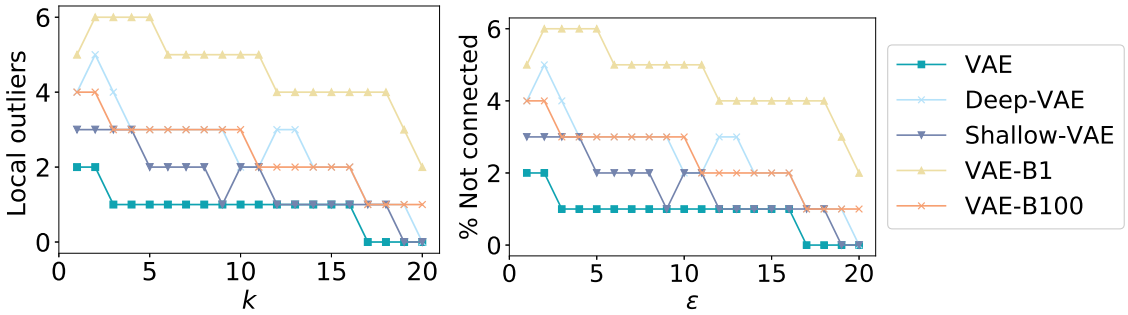


Figure 5.8: FCF group fairness scores when different VAEs are used. The green region represents the $[\sigma, \frac{1}{\sigma}]$ range indicating fair treatment.

Deep-VAE shows a slight over-fitting of the data and misses the discrimination in the SVM model that achieves a score of 1.12 (unfair treatment) with VAE and 1.02 (fair) with Deep-VAE. Shallow-VAE yields the same results as VAE. The size of the bottleneck layer shows a significant impact on the computation of FCF. A size of 1 is an under-representation of the relations in the data, and a size of 100 is an overfitting where the network memorizes the training data instead of learning meaningful relations. Both cases can reshape the neighborhood, based on the VAE latent representations, around particular instances yielding inaccurate fairness measures.

Furthermore, we study the faithfulness of FCF with different VAE architecture. To this end, we report the proximity and connectedness scores in Fig. 5.9. As demonstrated, faithfulness is significantly compromised with VAE-B1 and VAE-B100. Additionally, Shallow-VAE exhibits lower faithfulness scores than Deep-VAE. This experiment highlights the importance of a good approximation of the data distribution on faithfulness. It additionally shows the importance of the latent dimension in approximating the similarity metric implying a faithful comparison during fairness evaluation.



(a) Proximity (b) Connectedness

Figure 5.9: Impact of VAE architecture on faithfulness

5.5 Research Directions

We propose FCF, a *faithful* evaluation scheme to detect bias in black-box machine learning models. FCF leverages manifold distance metric computed by VAEs; thus, satisfies attainability constraints and can be extended to non-tabular datasets. Furthermore, FCF does not rely on the specification of distance or cost functions, which makes it easy for practitioners to adopt it in a wide variety of applications. We show that FCF can detect discrimination missed by other metrics that rely on input perturbation and causal theory.

We further propose *stability*, *proximity*, and *connectedness* as novel quantitative metrics to characterize fairness *faithfulness*. Our metrics show that FCF achieves

higher proximity while preserving stability and connectedness. Thus, FCF is faithful to the individuals by contrasting them to neighbors that are less likely to be outliers and unattainable. We quantitatively study the impact of the VAE complexity and the latent dimension on FCF scores and faithfulness.

FCF suggests several promising directions for future work. First, generative methods can be leveraged to obviate the need for training data while detecting bias. This can enhance privacy guarantees of our method complying with data laws and the GDPR. Second, a human evaluation study can be conducted to assess the relevance of the detected bias cases and further strengthen FCF.

Another promising avenue for future work includes the extension of FCF to unstructured datasets where sensitive attributes become harder to divulge. In the next objective, we devise an extension of FCF to textual data and we propose faithfulness evaluation methods. The application to computer vision settings will be studied in future work.

Chapter 6

Objective III: Extension of Contrastive Fairness and Faithfulness Evaluation to Textual Settings

After presenting our contrastive fairness evaluation for tabular data, we extend our evaluation technique to settings where the protected attribute is not explicitly manifested in the input. This is mainly the case in textual and imagery inputs. Given that the latter data type has recently been thoroughly studied in the work of [157], we focus this chapter on textual classifiers.

Despite their spectacular success in encoding semantic relationships, word embedding models can be contaminated by social discrimination [158]. This discrimination, in tandem with historical bias in training data, thwarts fair treatment for marginalized groups. For instance, recent studies have shown that sentiment analysis models are biased toward predicting negative emotions for people of color [159] and toxicity detection discriminates against homosexual individuals [48].

Unfortunately, the nature of textual data and the opacity of word embeddings

[65] impede the extension of existing fairness metrics [31], [38] to textual data. First, input perturbation in textual applications requires careful consideration in determining sensitive information and substituting corresponding words without violating grammatical rules [36], [53]. More importantly, implicit stereotypes learned by language models constitute a key challenge where embedding models can reflect genderial stereotypes even when gender-specific pronouns are omitted in training. Any bias detection approach that operates on word perturbations will then fall short of detecting the underlying stereotypes. Apart from measuring bias, there is a plethora of literature on debiasing embeddings [104], [124], [125], [160], but little has been done to address the bias in textual classifiers. De-biasing textual classifiers remains a challenge primarily when it currently relies on the availability and the quality of sensitive word associations [35], [36].

In this work, we examine in detail bias detection in text classification. Instead of perturbing the input $x \mapsto x'$ and observing the change in outcome, we consider the decision boundary between x and their contrastive counterparts x', i.e., inputs that can alter the model's outcome. Our **Co**ntrastive **F**airness **E**valuation (CoFE) framework deems a model f fair if the decision boundary between x and x' does not encode any sensitive information. CoFE measures bias through two metrics inspired by geometrical analogies and mutual information. Although the predominant SA studied in the literature is gender with some notable exceptions [35] to sexual orientation, CoFE broadens the scope to under-explored protected data such as *religion*.

Additionally, we consider the assessment schemes, and we target a novel evaluation aspect of the plausibility and attainability aspect of the textual contrastive examples. 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.

In what follows, we present the problem statement of our third dissertation objective in Section 6.1. We propose the methodology for fairness evaluation in textual classifiers in Section 6.2. We then describe our experimental design and empirical results in Section 6.3 and we evaluate the faithfulness of textual contrasts in Section 6.4 before discussing and concluding in Section 6.5.

6.1 Problem Statement

Given an input space $\mathcal{X}$, each $x \in \mathcal{X}$ is a sequence of words and an individual $x_i \in \mathcal{X}$. We assume a local factual world

$$\mathcal{F}_i = \{x | x \in \mathcal{N}_k(_i) \subset \mathcal{X}, f(x) = y_f\}$$

and its counterfactual world

$$\mathcal{C}_i = \{x' | x' \in \mathcal{N}_k(_i) \subset \mathcal{X}, f(x') = y_{nf}\}$$

, where $\mathcal{N}_k(_i)$ is the k-neighborhood of x_i computed according to the L_2-norm on the transformer encodings, y_f (y_{nf} resp.) is a favorable (non-favorable resp.) outcome. $|\mathcal{F}_i| + |\mathcal{C}_i| = k$ and k can be adjusted to allow sufficient training instances. Additionally, this diversity [161] is desirable as it enforces the comprehensiveness of our study.

We study the fairness of a predictor by studying whether the decision boundary between $\mathcal{F}$ and $\mathcal{C}$ encodes any sensitive attribute information. Furthermore, we study whether a given contrastive, in textual settings, is derived *faithfully* by extending the faithfulness metrics proposed in the previous chapter to the NLP framework.

6.2 Methodology

We operate in an IF framework and we study local neighborhoods where the decision-making can be rendered explainable via linear approximation. The neighborhood is sampled according to the transformer encoding of an input text. Then, a linear classifier is trained on the encodings. We then investigate the contrast in outcome between close inputs searching for any encoding of sensitive attributes. We do not perform any input perturbation, nor do we put any assumptions on the neighbors. Thus, we cover the implicit encoding of the sensitive attributes.

Accordingly, we define contrastive textual fairness as follows.

A textual classifier is fair to an individual x_i if the decision boundary between its factual to its counterfactual local neighborhoods does not encode sensitive information. Mainly, $(\mathcal{F}_i \rightsquigarrow \mathcal{C}_i) \perp\!\!\!\perp A$.

We approximate the decision boundary $\mathcal{F} \rightsquigarrow \mathcal{C}$ locally with a linear classifier while respecting local fidelity. Given x_i and k, the contrastive cost $\mathcal{F} \rightsquigarrow \mathcal{C}$ is calculated as the normal vector, $\overrightarrow{N}$, to the hyperplane separating between instances in $\mathcal{F}_i$ and $\mathcal{C}_i$. It is viewed as a local approximator of f and trained on labels produced by f that enforce the local fidelity constraint. The contrastive decision boundary is then studied along SA information to identify potential inter-dependence ($\perp\!\!\!\perp$) through geometrical and entropy-based approaches, as described in Figure **??**.

6.2.1 *Sensitive Attribute Information*

We consider the SA direction $\overrightarrow{a}$ to be constructed from pairs of related words. For example, the gender direction is $\overrightarrow{\texttt{man}} - \overrightarrow{\texttt{woman}}$, the age direction is $\overrightarrow{\texttt{young}} - \overrightarrow{\texttt{old}}$. Although this simple algebraic arithmetic is shown to capture diverse semantic relationships, we note the following. First, SA can be manifested in more than two categories (e.g. *race* can be white, African-American, Asian, etc.). Second, the

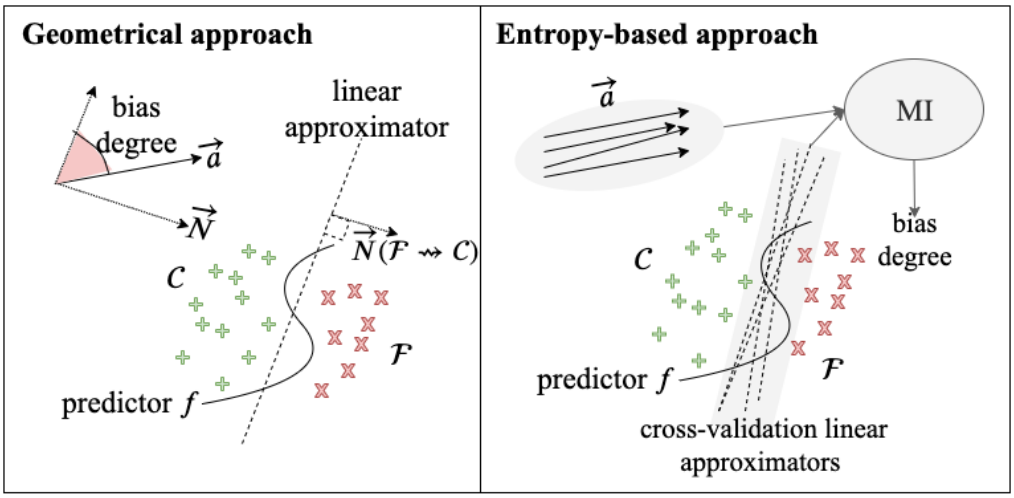


Figure 6.1: CoFE workflow

directions from different word combinations are not purely parallel. For instance, $\overrightarrow{\texttt{man}} - \overrightarrow{\texttt{woman}}$ might not be parallel to $\overrightarrow{\texttt{grandfather}} - \overrightarrow{\texttt{grandmother}}$.

Accordingly, we collect n pairs of (privileged, underprivileged) for each SA. These pairs can be non-overlapping (for example, for gender, (`male`, `female`), (`uncle`, `aunt`) or with some repetition for non-binary attributes (for example, (`American`, `Chinese`) and (`American`, `African`) for ethnicity. The SA direction can thus be considered as either (1) the direction that preserves the variance of the n samples through PCA, or (2) a multivariate random variable with n realizations.

6.2.2 *Sensitive Dependence Evaluation*

Now that we computed the SA direction, we devise two strategies to infer whether the decision boundary encodes sensitive data.

Geometrical Analogies

We compute the SA direction $\vec{a}$ similar to [104], by performing a PCA on the vector differences of the pairs (privileged, underprivileged). The PC that retains the majority of variance identifies the SA direction $\vec{a}$. An alternative formulation relies instead on a subspace formed by PCs M, but we restrict this study to $M = 1$,

which demonstrates effectiveness in detecting bias. We do not restrict our work to $k = 1$ in order to generalize our work to other protected attributes. The projection, w_P, of word w onto P is computed as:

$$w_P = \sum_{i=1}^{k} (w.b_i).b_i \tag{6.1}$$

Finally, the bias degree can be computed as the norm of the projection into the sensitive attribute subspace P or $||w_P||_2$. Accordingly, we measure the dependence as $cos(\overrightarrow{a}, \overrightarrow{N})$. We require the sensitive direction to be tangent to the decision boundary of f around x_i. Hence, bias is inferred by cosine scores < 1.

Mutual Information

Now, we reckon the decision boundary and the SA as random variables and we study the dependency through Mutual Information (MI). Given a cross-validation parameter C, we train C hyper-planes and we treat their coefficients as a d-dimensional random variable $\boldsymbol{U}$. Similarly, the SA is seen as a random variable $\boldsymbol{V}$ with each pair (privileged, underprivileged) as one realization. Our goal is to compute the amount of information that can be obtained from a $\boldsymbol{V}$ by observing $\boldsymbol{U}$ as $H(\boldsymbol{U}) + H(\boldsymbol{V}) - H(\boldsymbol{U}, \boldsymbol{V})$ where $H(.)$ refers to the Shannon entropy.

Luckily, there is a plethora of literature [162] on the estimation of the MI through sampling, popularized by the seminal work of [163]. These estimations rely on counting occurrences and co-occurrences [163], [164], kernel density estimations [165], [166], bounds on MI [167], [168]. In this work, we rely on MI based on $k-$nearest neighbors estimators[169] due to its simplicity and efficiency.

6.3 Results

Prior to our validation, we describe the datasets, the classifiers, and our implementation strategy. Then, we report the bias that existed in the dataset before any

training and then the bias leveraged by the different classifiers. Finally, we compare to existing fairness metrics in the literature.

6.3.1 *Experimental Setup*

We validate CoFE on two classification tasks: toxicity detection in comments [35] and occupation classification in biographies (bios) [170]. Fair toxicity detection ensures that minorities are not increasingly silenced by the moderation of comments on social networks. Bias-free occupation detection dilutes stereotypes in inferring identities from professions.

The classifiers in the toxicity prediction task are binary in nature and trained on 127,820 to rate toxic behavior. The favorable outcome is assumed to be *non-toxic* (that is, the toxicity rate is < 0.5). The occupation classifier is trained on $178,619$ examples to predict the occupation given the biography[1]. As the underlying classification is not binary, we create a list of stereotypical pairs that are well-known to exhibit discrimination. For instance, *surgeon* is favorable while *nurse* is not. Similar pairs are *attorney-paralegal*, *physician-dietitian* and *professor-teacher*, *personal trainer-model*.

We experiment with four underlying transformers adopted from [171]: BERT, RoBERTa, XLM, GPT[2] and a Convolutional Neural Network (CNN) with a hidden size of 128 and a kernel size of 5 followed by a fully-connected layer of 10 units trained on GLoVe embeddings [172]. CoFE is evaluated on the validation data of both datasets on an NVIDIA K80 / T4 GPU with 16GB RAM. Experiments were repeated 5 times and average results are reported.

[1]The data is available at github.com/microsoft/biosbias

[2]'unitary/toxic-bert', 'unitary/unbiased-toxic-roberta', 'unitary/multilingual-toxic-xlm-roberta' and 'martin-ha/toxic-comment-model' on `huggingface` respectively.

6.3.2 *Computation of the Sensitive Direction*

We consider SAs that might emerge in textual data, i.e. *age*, *race*, *gender*, *religion*, and *sexual orientation*. The SA direction is based on gender word pairs of [104] and on home-grown collection for other SAs. We show five random pairs in Table 6.1 for each SA. Additionally, we consider a direction derived from randomly selected words from the English vocabulary with no clear analogies as a baseline for our study.

Baseline	*(tooth, committee), (cabinet, lab), (instance, son), (champion, recording), (family, heart)*
Age	*(young, old), (youthful, elderly), (young, elderly), (youthful, old), (young, senior)*
Race	*(white, indian), (white, korean), (american, dark skin), (european, african-american), (white, black)*
Religion	*(christian, muslim) (Jesus, Muhammad), (church, mosque), (orthodox, muslim), (baptism, Hajj)*
Sexual orientation	*(straight, homo), (straight, bisexual), (man, gay), (straight, gay), (straight, homosexual), (straight, homophile), (straight, queer)*
Gender	*(husband, wife), (uncle, aunt), (brother, sister), (dad, mom), (waiter, waitress)*

Table 6.1: Pairs of sensitive attributes

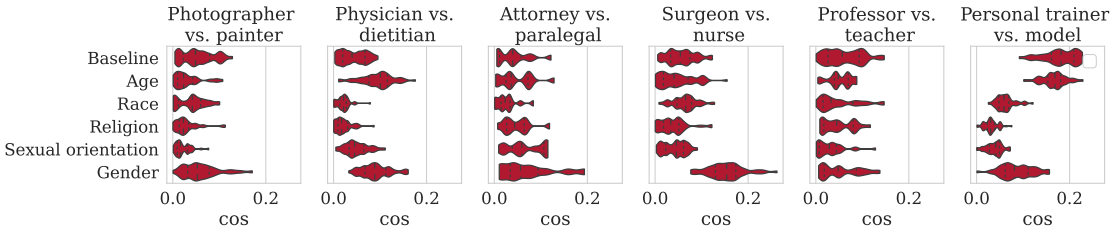


Figure 6.2: CoFE bias scores in the bios dataset

6.3.3 *Implementation Details*

Sentence embeddings are computed as 768-dimensional dense vectors based on the recent work of [173]. No significant impact was noticed with different embedding models.

	BERT			
	CoFE % (↓)	EqOpp (↓)	EqOdds (↓)	Acc. (↑)
Toxicity (Race)	0.333	0.154	0.183	0.881
Toxicity (Religion)	0.41	0.02	0.008	0.881
Toxicity (Sexual Orientation)	0.466	0.584	0.721	0.881
Toxicity (Gender)	0.48	0.327	0.765	0.881
Bios (Surgeon vs. nurse)	0.398	0.427	0.56	0.972
Bios (Lawyer vs. Paralegal)	0.346	0.321	0.764	0.991
	CNN			
	CoFE % (↓)	EqOpp (↓)	EqOdds (↓)	Acc. (↑)
Toxicity (Race)	0.41	0.167	0.169	0.714
Toxicity (Religion)	0.461	0.018	0.135	0.714
Toxicity (Sexual Orientation)	0.692	0.433	0.669	0.714
Toxicity (Gender)	0.641	0.267	0.612	0.714
Bios (Surgeon vs. nurse)	0.399	0.325	0.456	0.875
Bios (Lawyer vs. Paralegal)	0.256	0.321	0.54	0.813

Table 6.2: CoFE bias vs. equality of opportunity and equalized odds metrics

For local approximation, we sample $k = 100$ neighbors based on the cosine similarity of the latent representations. Two factors contributed to the choice of k: (1) k needs to be large enough to yield a neighborhood where favorable and unfavorable outcomes appear. (2) k should not exceed $\approx 10\%$ of the test data for the linear model to faithfully approximate f. In order for CoFE to satisfy large-scale retrieval requirements, we rely on the anisotropic vector quantization of the ScaNN algorithm [174] to sample $k-$nearest neighbors. Based on ScaNN heuristics, we select the dot product similarity metric with the following hyper-parameters for the search strategy: *num_leaves* $= 200$, *num_leaves_to_search* $= 100$, *training_sample_size* $= 250$ and an *anisotropic_quantization_threshold* of 0.2 for scoring. Adopting ScaNN reduced the local sampling time from hours needed by brute-force similarity algorithms to a few minutes on our hardware.

When we are interested in the bias cases, we display the inverse of CoFE cosine $(1 - \cos(.))$ and call it CoFE bias (0 means no bias, 1 means severe bias). When we are interested in the discrete number of bias cases, we display CoFE%, which denotes the percentage of individuals whose CoFE bias is above a threshold th. It should

be noted that CoFE% establishes the group fairness counterpart of our individual score, CoFE. Furthermore, we omit MI scores and report cosine scores only when both metrics contribute to the same conclusion.

6.3.4 *Dataset Bias*

Prior to any training, we study the bias in the dataset by sampling 50 random points and computing CoFE bias on the true labels. Figures 6.2 and 6.3 show the CoFE bias scores in the considered datasets. Discrimination manifests itself when higher CoFE scores are observed with regard to the baseline distribution. Consequently, for the bios data, only *gender* exhibits a noticeable bias. This discrimination is mainly shown in the *doctor-nurse* and *attorney-paralegal* contrast. In the rest of this study, we will only consider *gender* as SA in the two aforementioned categories. For toxicity data, a noticeable bias is observed against *race*, *religion*, *gender*, and *sexual orientation* is observed. The CoFE scores on the age attributes are even lower than the baseline suggesting fair treatment for younger and elder groups. Therefore, age will not be a subject of this study.

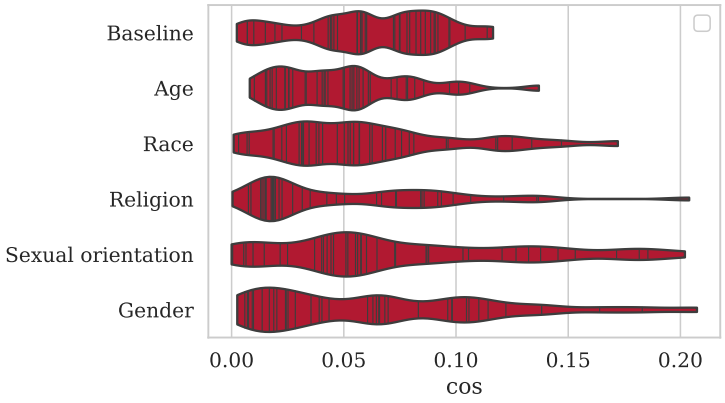


Figure 6.3: CoFE bias scores in toxicity dataset

In what follows, we report the detected bias in the two datasets and we compare it to existing metrics.

6.3.5 *Detected Bias*

For both datasets, we sample 100 random points where the outcome is not favorable and compute their CoFE% cosine score. The threshold *th* is chosen according to the results of the baseline direction where a model can exhibit CoFE scores of up to 0.1 even when no bias is discerned. Furthermore, we compute the equality of opportunity (EqOpp) and equalized odds (EqOdds) scores as in [36]. The former is the difference between true positive (TP) rates in both groups (0 <EqOpp < 1) the latter is the difference between TP and the false positive (FP) rates (0 <EqOdds < 2). A zero value of EqOdds and EqOpp is favored, as it hints at fair treatment. EqOpp is an appropriate fairness evaluation metric in the Bios data. It ensures that privileged and underprivileged groups are equally matched with jobs that are relevant to their skills. EqOdds is suitable in toxicity prediction settings to ensure equal odds of detecting toxicity in comments with equal rates across all groups.

We report EqOdds and EqOpp along with CoFE% scores in Table 6.2 and we show the distribution of the cosine scores in Figure 6.4. *Gender* and *sexual orientation* manifest notably higher bias (score 0.7 with BERT) on the toxicity dataset. This score significantly exceeds the original bias of the dataset (score 0.2). Occasionally, BERT yields higher discrimination according to all metrics which can be explained by the historical bias perpetuation during pre-training. Interestingly, one can see how CoFE detects bias cases that are diluted, even missed, by EqOpp, especially with regard to *religion*.

6.3.6 *CoFE in the Fairness Evaluation Realm*

Quantitative comparison to existing work such as [35], [36], [53], [175] is not viable due to the unavailability of their implementations which hinders reproducibility. We spare no effort at conducting a qualitative and analytical comparison. To show how word-level evaluations might miss discrimination cases, we compute CoFE bias

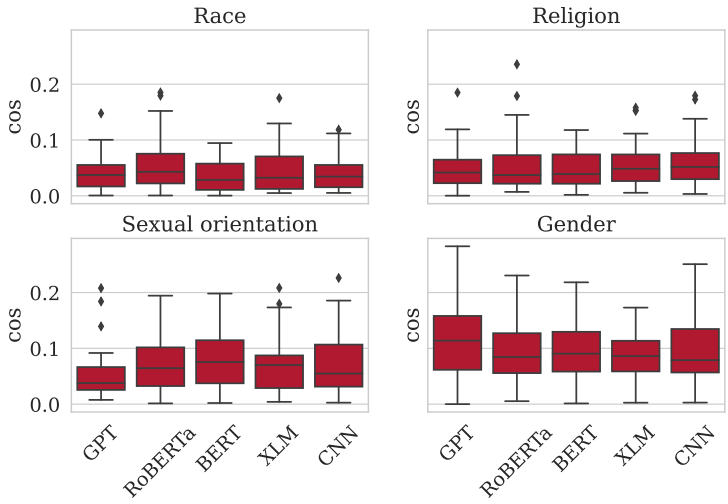


Figure 6.4: CoFE bias cosine scores on the toxicity dataset

scores on biographies by eliminating gender-specific words and pronouns. Figure 6.5 shows how CoFE bias cos and MI scores can show discrimination, especially in the surgeon-nurse stereotype which motivates the use of metrics in the latent space.

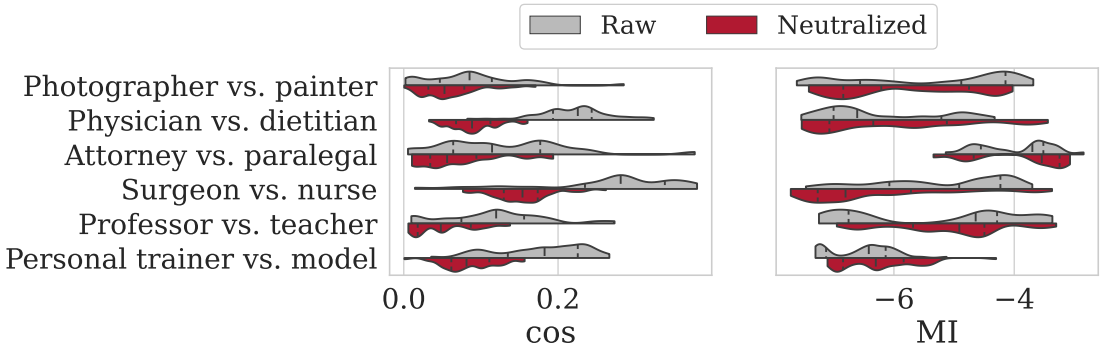


Figure 6.5: CoFE bias on bios data when raw and gender-neutralized biographies are used for training

Finally, scaling CoFE to languages with grammatical gender such as Spanish does not face the same challenges as in word-level methods [36]. As long as language models are reliable, CoFE only requires a handful of (privileged, underprivileged) pairs for evaluating textual classifiers in new languages. Additionally, CoFE is not limited to gender or binary discrimination/classifiers. It relies on privileged-underprivileged information which is suitable even for non-binary attributes (e.g.

race or *sexuality*).

6.4 Faithfulness of Textual Contrasts

In this section, we consider contrastive textual examples and we study their faithfulness based on the metrics we proposed earlier. We consider textual contrastive explanations 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 [3] with 977 validation instances.

For a given input, POYJUICE's API generates 3 counterfactuals whereas MiCE generates a variable number of counterfactuals between 1 and 69 as shown in Figure 6.6.

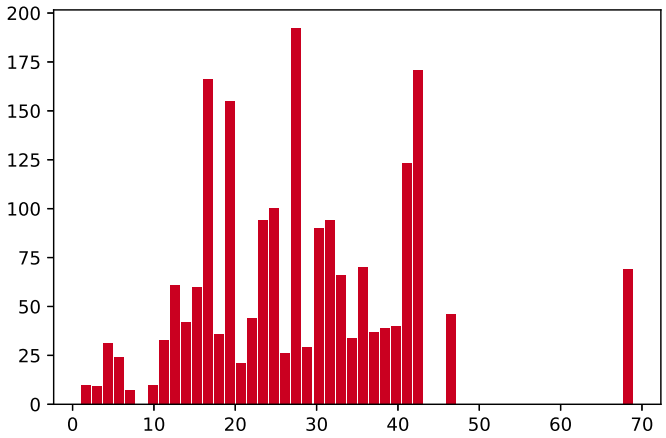
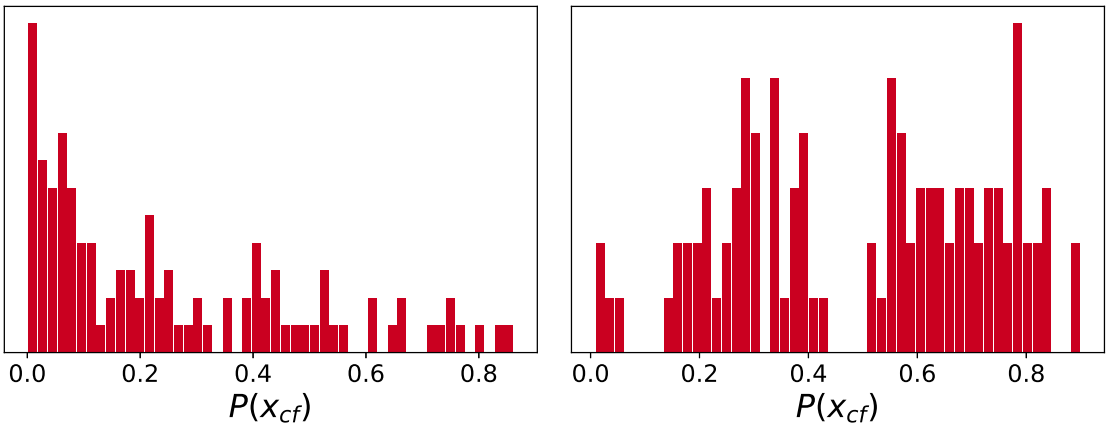


Figure 6.6: Distribution of the number of counterfactuals generated by MiCE for each input

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

6.4.1 *Proximity*

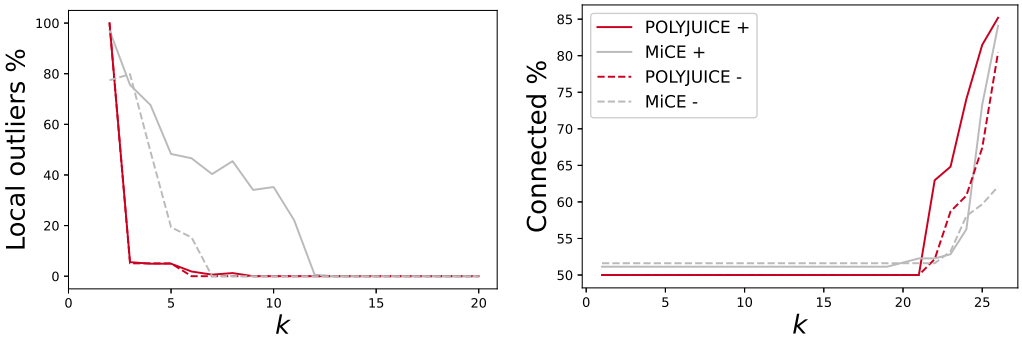
We start by evaluating how close the generated counterfactuals are to ground truth data from the contrast class. For this purpose, we compute $P(\boldsymbol{x}_{cf})$ and plot the distribution of its values in Figure 6.7. One can see a predominance of low proximity scores (< 0.2) in POLYJUICE and an inclination to achieve higher scores with MiCE.



(a) POLYJUICE (b) MiCE

Figure 6.7: Distribution of the $P(\boldsymbol{x}_{cf})$ scores

We further split our validation data according to their contrast classes into two categories: positive and negative sentiment contrasts. 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 6.8a. 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 contrast class being positive or negative sentiments.



(a) Proximity scores (b) Connectedness scores

Figure 6.8: Scores while changing the number of neighbors k

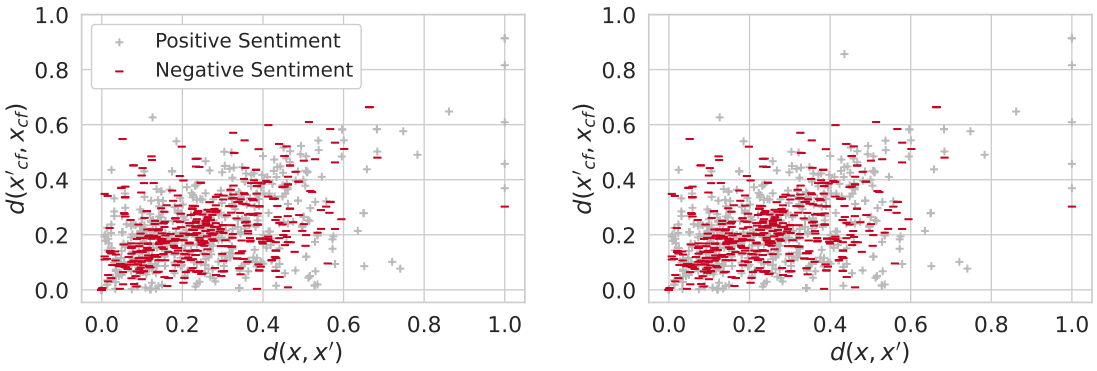
6.4.2 *Connectedness*

To assess whether the generated counterfactuals are connected to their original factuals, we compute the connectedness score for both explanation models and contrast sentiments. The results shown in Figure 6.8b 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 contrast classes seem to achieve higher connectedness scores, but the discrepancy between positive and negative sentiments is insignificant with MiCE.

6.4.3 *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 6.9 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 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 contrast categories.



(a) POLYJUICE (b) MiCE

Figure 6.9: Scattering of counterfactual similarity with respect to the input similarity. Linear scattering infers local stability.

Finally, we consider a more fine-grained stability study, 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 6.10 shows how the counterfactual similarity is distributed for the three ranges considered. Locally, i.e., with input distance < 0.2, POLYJUICE is shown to be more stable in the positive contrast class by achieving low distances in the generated counterfactual. MiCE seems to outperform POLYJUICE on the negative contrast class. Zooming out, better stability is observed with POLYJUICE for both contrast classes.

6.4.4 *Adversarial Robustness*

We generate adversarial perturbations based on semantic similarity [112] on the restaurant reviews. The adversarial inputs are then fed into POLYJUICE and MiCE for a counterfactual generation. Figure 6.11a 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

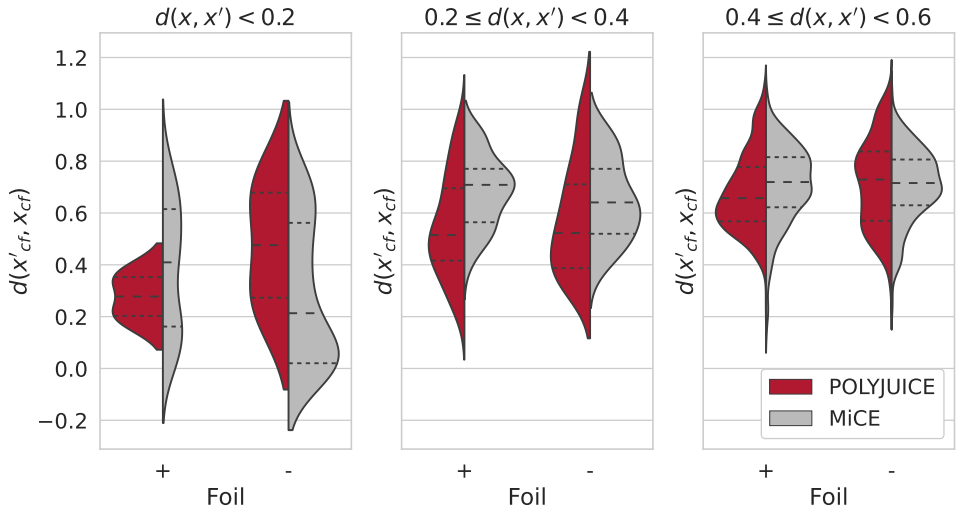
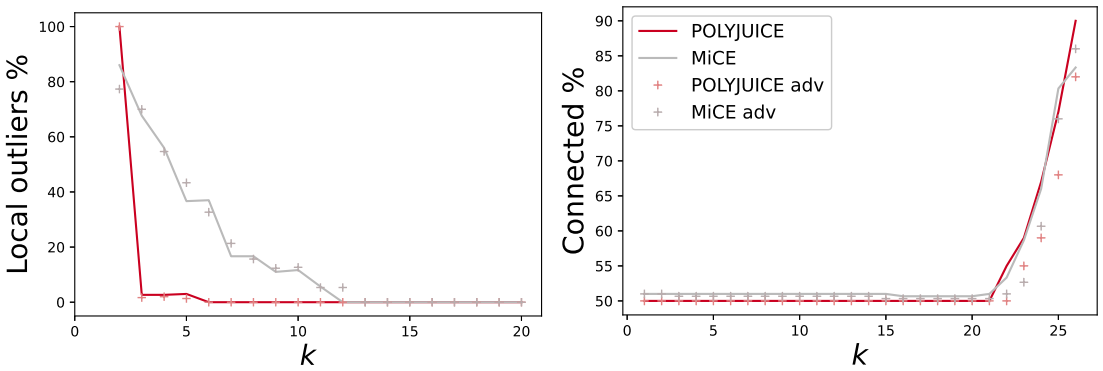


Figure 6.10: Distribution of the distance between counterfactuals for different input distance ranges

robustness to adversarial attacks. The connectedness scores are not affected for both methods, as shown in Figure 6.11b.



(a) Proximity scores

(b) Connectedness scores

Figure 6.11: Proximity and connectedness results with adversarial attacks on textual contrastive examples

Finally, we visualize how the generated counterfactuals are affected when inputs are perturbed. Figure 6.12 shows the distribution of 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 counterfactu-

als showing better robustness to adversarial attacks. Since POLYJUICE does not rely on gradient descent to reach recourse, its results are per the discussion of [176] 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.

(a) POLYJUICE (b) MiCE

Figure 6.12: Distribution of the cosine similarity of the generated counterfactuals with adversarial attacks

6.4.5 *Comparison to Existing Metrics*

We compute the 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.

The results show a slight improvement of POLYJUICE over MiCE which confirms our findings highlighting again the importance of latent representations. Figure 6.13 shows the distribution of the scores on the counterfactuals generated by POLYJUICE and MiCE.

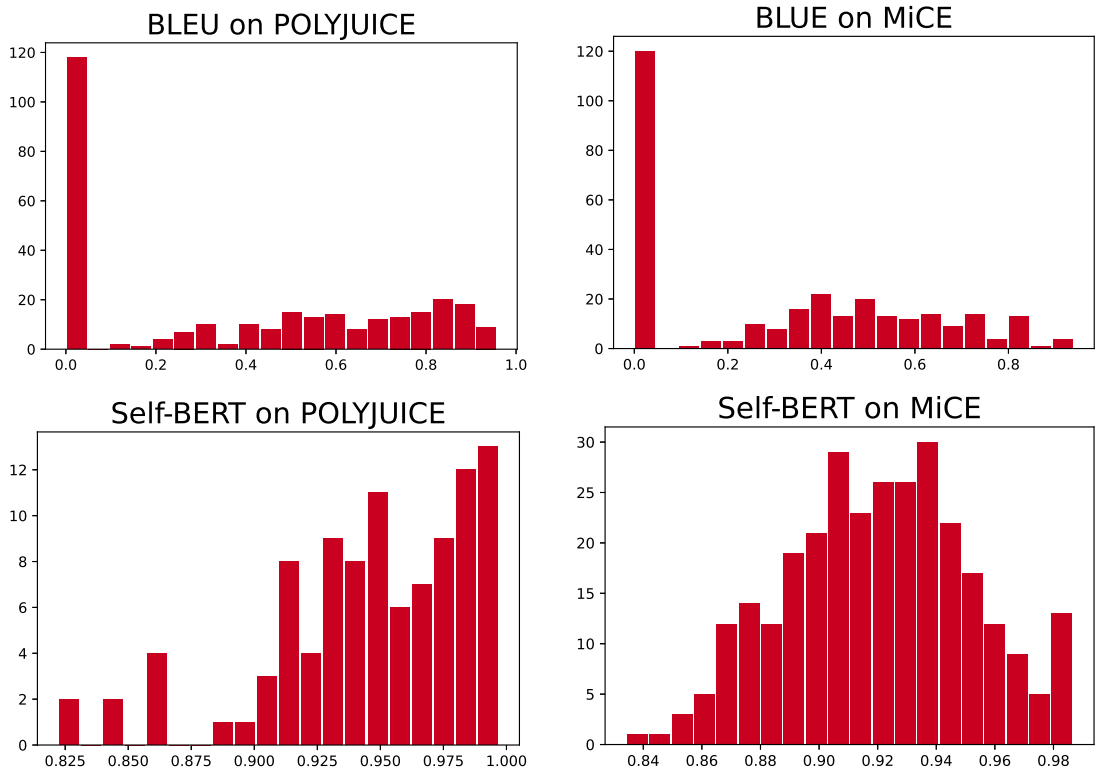


Figure 6.13: Distribution of the BLEU and Self-BERT scores on the generated counterfactual textual examples

6.4.6 *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 lens of the word representation.

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.

On the contrary, the connectedness scores do not show any substantial difference. POLYJUICE has been 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. Additionally, latent representations of words are shown to be more reliable with semantic adversarial attacks.

Finally, *faithfulness* is not shown to be distorted towards one sentiment versus the other. The consistency in the results reported on the positive and negative sentiment suggests a balanced training strategy.

6.5 Research Directions

This work adds to the growing body of fairness research by suggesting CoFE, a novel contrastive evaluation technique for textual classifiers. By addressing NLP challenges and leveraging transformers' training, CoFE exposes bias and stereotypes learned by textual classifiers that are diluted, even missed, by other fairness metrics. CoFE considers a variety of sensitive attributes including *religion* and *homosexuality* and is robust to any deep architecture.

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.

A limitation of CoFE is its sensitivity to some definitions where a higher bias is observed with words with little consideration for the context. Diluting such sen-

sitivity can be the subject of future work. Another promising line of work is the extension of CoFE to languages with different linguistic properties such as Spanish.

We suggest three immediate steps in the textual faithfulness line of work. First, "unconnected" counterfactuals should be filtered by posing connectedness constraints on the search strategy. Second, stability should be enhanced when textual edits are employed. Finally, textual attacks on recourse methods can be further investigated to propose robustness methods and mitigation techniques.

Chapter 7

Objective IV: Latent Bias Mitigation

After presenting how contrastive learning can be used to derive counterfactual examples and detect bias cases in classifiers, in this chapter, we present bias mitigation through contrast sets. We devise two techniques that rely on contrastive sets to force our contrastive fairness on any classifier.

Instead of relying on adversarial learning and re-weighting schemes [128]–[132], we propose a regularizer that is aligned with our evaluation strategy. To this end, we suggest a mitigation technique that neutralizes a classifier by augmenting its loss function with a contrastive fairness constraint. Our constraint encourages a classifier to treat proximate individuals similarly while considering a manifold-like notion of distance.

Then, we extend this technique to textual settings to suggest a latent augmentation technique with no assumptions on the underlying text classifier. Existing work relies on pre-defined word analogies for gender mostly [52], [104]. We highlight two aspects that prevent the wider adoption of word-based augmentation approaches. Their first maneuver is to establish whether a word is related to a potentially biased direction which is prone to error. Second, they require a significant human

intervention step to identify privileged-underprivileged analogies in text prior to augmentation. Accordingly, we suggest the first augmentation technique that operates on the latent rather than the input space of classifiers. This eliminates the reliance on sensitive analogies as a preparatory manual step. We study its effectiveness in de-biasing textual classifiers and the impact of the mitigation strategy on the model's performance.

This chapter presents the fourth dissertation objective as follows. First, we define our bias mitigation problem statement in Section 7.1 and we discuss our methodology in Section 7.2. Then, we present our empirical results in Section 7.3 and we discuss promising future directions in Section 7.4.

7.1 Problem Statement

Given a predictor $\hat{f}$, an individual x_i, and a contrastive example x_j derived based on a manifold-like distance, we encourage $\hat{f}$ to treat x_i and x_j similarly.

x_i can be a numerical record or a textual example represented by an embedding vector. In the former case, we assume x_j to be derived according to the methodology discussed in Algorithm 1 in Section 5.2. In the textual settings, we derive x_j by perturbing x_i with the sensitive attribute direction. We rely on the methodology described in Section 6.2.1 to compute this direction.

7.2 Methodology

We suggest two mitigation strategies. The first one regularizes the loss of a classifier with a contrastive fairness constraint. This technique is applicable in general settings as an in-training step. Our second strategy augments the datasets with latent contrastive examples. While it can be extended to different datatypes, we discuss our second technique in NLP, within the same settings as our evaluation technique.

7.2.1 *Regularizationn in General Settings*

We propose a faithful contrastive regularizer that forces a classifier to generate the same prediction for an individual x_i and its contrastive neighbors $\mathcal{N}_\epsilon^{CT}(x_i)$. The regularizer augments the loss function of a classifier $\hat{f}$, $\mathcal{L}(\hat{f}(x_i), y_i)$, with a constraint that the predictions of an individual and its contrast are the same, i.e.

$$\mathcal{L}(\hat{f}(x_i), y_i) + \lambda MSE(\hat{f}(x_i), \hat{f}(x_j))$$

with y_i is the ground truth label for x_i, λ is a regularization hyper-parameter, MSE is the mean-squared error and $x_j \in \mathcal{N}_\epsilon^{CT}(x_i)$.

Classifiers that do not optimize a loss function, such as decision trees, can benefit from our FCF formulation through contrastive augmentation. In other words, the training dataset can be augmented with tuples (x_j, y_i) where $x_j \in \mathcal{N}_\epsilon^{CT}(x_i)$ and y_i is the label of x_i. Both techniques can be used as a post-hoc debiasing technique of pretrained models and as an inherent way to enforce contrastive fairness constraints.

7.2.2 *Latent Augmentation in Textual Settings*

In [36], a model f is *certifiable robust* in the context of fairness if, for any sentence x and its alternatives x', $f(x) = f(x')$. x and x' carry the same meaning in the classification context but differ in SA. This condition is enforced through augmentation on the word and sentence levels [52], [53]. We suggest latent augmentation of the input encoding. This technique alleviates the need for an SA topology and for complete substitution databases.

Accordingly, for a dataset $\{(x_i, y_i)\}$, we augment with $\{(x_i \pm \overrightarrow{a}, y_i)\}$. With $\overrightarrow{a}$ computed as in 6.2.1. Our augmentation encourages the model to have a similar treatment for individuals that carry the same meaning except for the SA. It enhances robustness against *semantic biased attacks* to achieve fairness while preserving rich

semantics. Additionally, our augmentation does not require pre-training as the models can be only fine-tuned on the augmented data or part of hereby.

We draw the reader's attention to the distinction between our work and the debiasing of [104], [124], [125], [160], [177]. The latter approaches eliminate implicit stereotypes from embedding models. CoFE, however, requires a textual classifier to reach the same decision for two similar inputs from different protected groups whether the SA is implicit (stereotypical) or explicit (historical classification bias).

7.3 Results

In what follows, we cover the experimental setup of both mitigation techniques. Then, we report the results of our debiasing strategies in general and textual settings. Furthermore, we study the impact of our mitigation technique on the fairness-accuracy trade-off and we compare it to existing work.

7.3.1 *Experimental Setup*

For the general settings, we follow the same setup described in Section 5.4. Mainly, we test our metric on the adult census dataset, used to predict whether an adult's income $> 50K$ USD where race and gender are sensitive attributes. We use 60% of each dataset to train the model, 20% to validate the model and select the hyper-parameters, and 20% to test the FCF scores. We experiment with a Decision Tree (DT), Support Vector Machines (SVM) model, Logistic Regressor (LR), and an Artificial Neural Network (ANN). DT is trained on Gini impurity with a maximum depth of 5 and SVM is trained with an RBF kernel. LR is an $l2$-penalized model and we consider $d_2 = 5$. ANN consists of 2 hidden layers of 13 and 4 neurons activated via ReLU and trained using a weighted binary cross-entropy loss function through gradient-descent with root mean squared propagation for 50 epochs with a 10^{-3} learning rate.

For the textual classifiers, we employ the two aforementioned datasets: toxicity detection in comments [35] and occupation classification in biographies (bios) [170]. We follow the same setup of Section 7.3 where we train the classifier of the toxicity prediction task on 127,820 to rate the toxic behavior. and the classifier of the occupation on $178,619$ examples to predict the occupation given the biography. We used the list of stereotypical pairs introduced in Section 7.3 and we experiment with BERT a Convolutional Neural Network (CNN) with a hidden size of 128 and a kernel size of 5 followed by a fully-connected layer of 10 units trained on GLoVe embeddings [172]. CoFE is evaluated on the validation data of both datasets on NVIDIA K80/T4 GPU with 16GB RAM. Experiments were repeated 5 times and average results are reported.

7.3.2 *General Settings*

We modify the loss function of the ANN model and we augment the training of DT, SVM, and LR with contrastive examples. We report the FCF group fairness scores along with existing metrics on the Adult dataset pre- and post-debiasing in Table 7.1. Additionally, we report the individual FCF scores on the 3 datasets in Figure 7.1.

	Debiased	**Acc**	**EqOpp**	**PredP**	**PredE**	**AccEq**	**StatP**	**FCF**
DT	No	84	1.00	1.02	0.99	0.98	0.99	0.91
DT	Yes	85	1.00	1.00	0.99	0.99	0.99	0.99
SVM	No	85	1.02	0.98	1.10	1.01	1.00	1.12
SVM	Yes	84	1.03	0.99	1.00	1.01	1.00	1.02
LR	No	83	0.89	1.22	1.02	0.97	1.02	1.23
LR	Yes	85	0.92	1.03	1.00	0.99	1.01	0.99
ANN	No	86	0.91	1.13	1.02	1.10	1.03	1.20
ANN	Yes	86	0.99	1.01	1.02	0.98	1.01	0.99

Table 7.1: Group fairness metrics along with FCF on the Adult dataset where underlined scores indicate discrimination with $\sigma = 0.9$

Our debiasing strategy is shown to result in high FCF scores (> 0.95) implying

fair treatment for individuals. Additionally, as shown in Table 7.1, the proposed mitigation strategy is shown to improve the fairness of different models based on our contrastive notion as well as other group fairness definitions. The bias cases reported earlier (with SVM, LR, and ANN) were significantly reduced without compromising the model's performance. On the contrary, the model's accuracy is improved in the case of DT and LR which can be attributed to the data augmentation that we followed. It is worth mentioning that no significant impact on the faithfulness scores was observed.

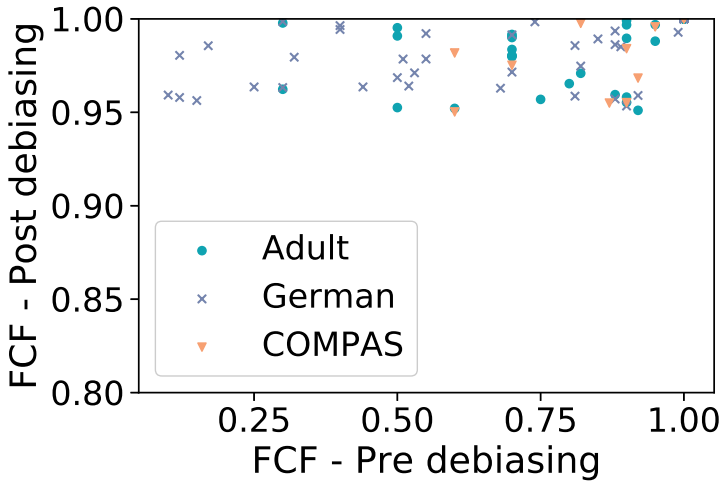
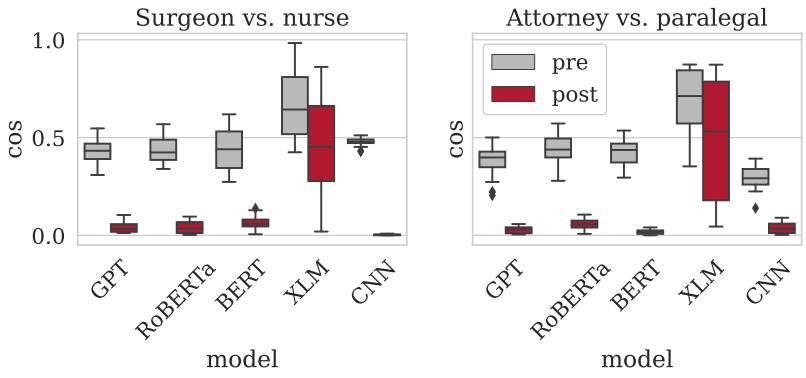


Figure 7.1: FCF pre- and post- bias mitigation

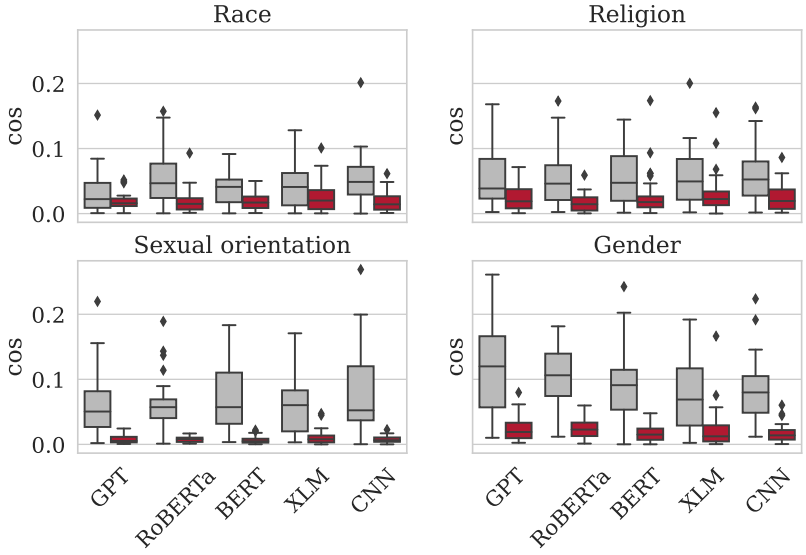
7.3.3 *Textual Debiasing*

For improved robustness, we apply a scalar transformation on $\overrightarrow{a}$ of magnitude α with $-1 \leq \alpha \leq 1$, a random number with uniform distribution. Mainly, for a tuple (x_i, y_i), we augment the dataset with $\{(x_i + \alpha\overrightarrow{a}, y_i)\}$ with a probability m. $\alpha < 0$ accounts for perturbations towards the underprivileged group.

Figures 7.2 and 7.3 report the distribution of the CoFE bias cosine scores and MI scores for the bios and toxicity dataset pre- and post-mitigation. The results demonstrate a significant improvement in fairness with latent mitigation where cosine bias scores are reduced by more that 50% for XLM and dropped to near zero



(a) Bios

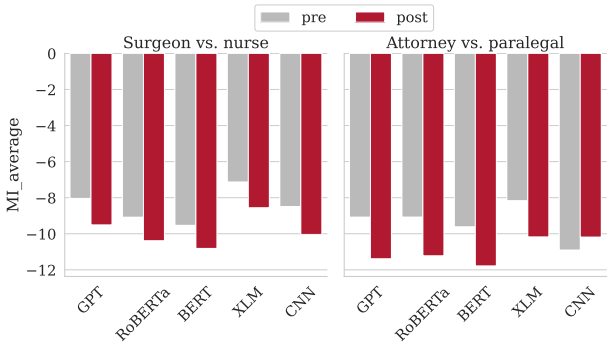


(b) Toxicity

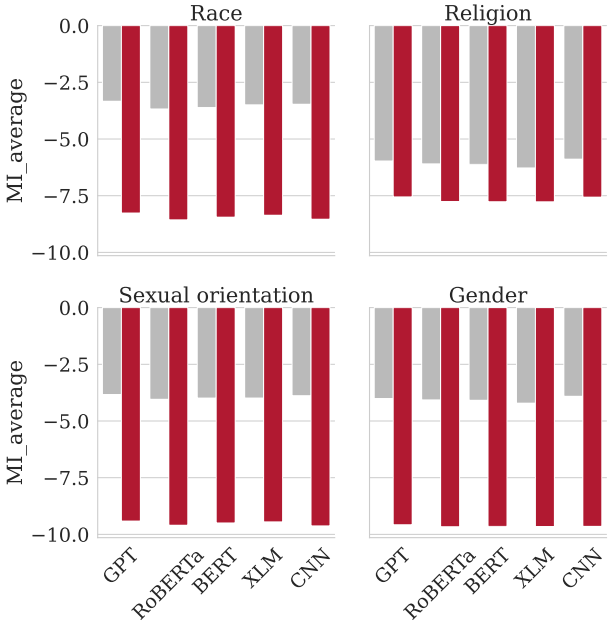
Figure 7.2: CoFE bias cosine scores for pre- and post-debiasing where higher cosine scores infer discrimination)

values with in other architectures. MI results show a similar improvement with up to 3x decrease especially with *sexual orientation*.

Table 7.2 further indicates a consistent improvement of up to 50% in EqOpp and EqOdds with different SAs in BERT and CNN. The example in Figure 7.4 demonstrates that latent augmentation removes correlations between `her` and *nurse* and corrects the correlation between `clinical` and the true occupation. While `her`-*nurse* correlation can be reduced with word augmentation, `clinical` might exhibit high correlation with *male* and *surgeon* that can't be remedied by word



(a) Bios



(b) Toxicity

Figure 7.3: CoFE MI scores for pre- and post-debiasing (Higher MI indicates discrimination)

augmentation given that no *male-female* analogy exists for `clinical`.

7.3.4 *Fairness-Accuracy Trade-off*

It is of utmost criticality for any fairness enforcement approach to preserve the model's performance quality. Table 7.2 shows that with BERT, the reduction in accuracy does not exceed 3%. The case of *sexual orientation* even shows an improvement in the accuracy. CNN's results are similar with a slightly higher drop in accuracy to up to 4%.

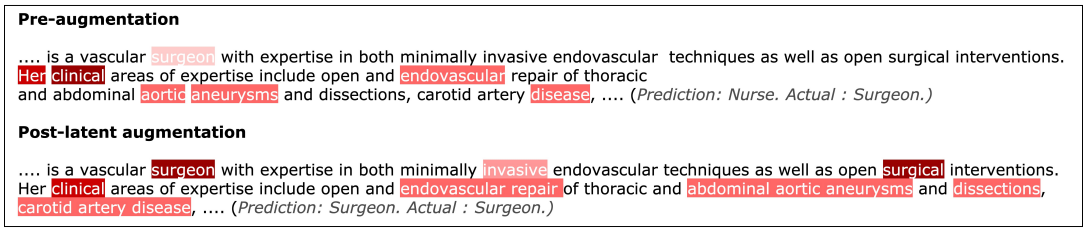


Figure 7.4: Example from bios data with LIME explanations highlighted in red (intense means higher correlations)

To better understand the accuracy-fairness tension with CoFE, we change the augmentation ratio m from 0 (no augmentation) to 1 (full augmentation) on the training data of both datasets and we improve the change in accuracy and CoFE cos scores. Figure 7.5 shows the average accuracy and CoFE bias scores of BERT and CNN on gender discrimination. One can see that augmentation improves fairness without compromising accuracy. More importantly, with high augmentation ratios, the performance can improve suggesting that our augmentation is treated as a semantic one and addresses over-fitting. Interestingly, fairness improvement is capped at $m = 0.4$ which can be used as a relaxation of the full augmentation requirement for efficiency and scalability.

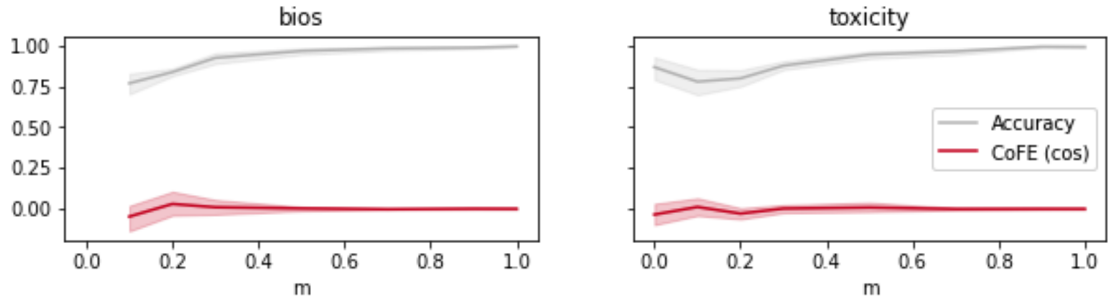


Figure 7.5: Accuracy and CoFE with augmentation ratio

7.3.5 *Comparison to Existing Work*

Unlike [35], [36], [52], [53], [175], CoFE does not depend on substitution databases and does not require the identification of sensitive words as a preliminary step. Additionally, the augmentation in [36] improves EqOdds and EqOpp by up 20% on

the bios and toxicity datasets. CoFE latent augmentation was able to drop these bias metrics by up to 50% in the gender case in the toxicity dataset. Similarly, [35] reduced EqOpp toxicity data with regard to homosexuals by 20% while CoFE reduces the same metrics by up to 30%. Additionally, word substitution methods [36], [178] were reported to require 53 hours for toxicity and 37 hours for bios data on large AWS compute nodes. CoFE augmentation exhibits a reduced carbon footprint when running on two Intel(R) Xeon(R) CPU cores of 2.30GHz and 12 GB RAM while considering 4 sensitive attributes instead of sexual orientation only.

Finally, scaling CoFE to languages with grammatical gender such as Spanish does not face the same challenges as in word-level methods [36]. As long as language models are reliable, CoFE only requires a handful of (privileged, underprivileged) pairs for evaluating and mitigating textual classifiers in new languages. Additionally, CoFE is not limited to gender or binary discrimination/classifiers. It relies on privileged-underprivileged information which is suitable even for non-binary attributes (e.g. *race* or *sexuality*).

7.4 Research Directions

In this chapter, we suggested mitigation techniques for classifiers that operate on tabular and textual data. Our strategy has been shown to be effective in neutralizing different classifiers with respect to different sensitive attributes. Additionally, our textual latent mitigation encourages fair treatment and improves robustness without compromising performance.

Our bias mitigation strategy breathes new flexibility towards (1) a painless extension to new languages and sensitive attributes and (2) debiasing classifiers infected by stereotypes rather than social bias in training data. A future direction is an additional improvement of the quality of the derived textual contrasts by filtering based on scores such as BLEU and Self-BERT scores.

	BERT							
	CoFE % (↓)		Equality of opp. (↓)		Equalized odds (↓)		Accuracy (↑)	
	None	Latent	None	Latent	None	Latent	None	Latent
Toxicity (Race)	0.333	0.002	0.154	0.059	0.183	0.071	0.881	0.896
Toxicity (Religion)	0.410	0.001	0.020	0.019	0.008	0.002	0.881	0.890
Toxicity (Sexual Orientation)	0.466	0.001	0.584	0.274	0.721	0.302	0.881	0.916
Toxicity (Gender)	0.480	0.001	0.327	0.104	0.765	0.247	0.881	0.842
Bios (Surgeon vs. nurse)	0.398	0.001	0.427	0.222	0.560	0.270	0.972	0.973
Bios (Lawyer vs. Paralegal)	0.346	0.005	0.321	0.147	0.764	0.153	0.991	0.980
	CNN							
	CoFE % (↓)		Equality of opp. (↓)		Equalized odds (↓)		Accuracy (↑)	
	None	Latent	None	Latent	None	Latent	None	Latent
Toxicity (Race)	0.410	0.002	0.167	0.070	0.169	0.067	0.714	0.671
Toxicity (Religion)	0.461	0.003	0.018	0.007	0.135	0.010	0.714	0.692
Toxicity (Sexual Orientation)	0.692	0.001	0.433	0.127	0.669	0.450	0.714	0.694
Toxicity (Gender)	0.641	0.002	0.267	0.031	0.612	0.109	0.714	0.686
Bios (Surgeon vs. nurse)	0.399	0.005	0.325	0.145	0.456	0.229.	0.875	0.872
Bios (Lawyer vs. Paralegal)	0.256	0.001	0.321	0.142	0.540	0.150	0.813	0.872

Table 7.2: CoFE bias vs. equality of opportunity and equalized odds metrics pre- and post- latent mitigation

Chapter 8

Conclusion and Future Directions

This dissertation aims at providing a comprehensive framework for explainable and fair AI that is inspired by contrastive learning, goes beyond tabular data through deep feature inspection, and satisfies faithfulness guarantees. Accordingly, we addressed the following research questions:

1. How to derive contrastive examples in the context of explainable AI while simultaneously accounting for immutability, semi-immutability, and attainability constraints in a model-agnostic fashion?

2. How does contrastive learning, with manifold-like distance measures, improve individual fairness evaluation to faithfully detect bias in classifiers?

3. How to utilize deep feature inspection to extend the contrastive fairness measure to non-tabular data?

4. How to leverage the derived contrastive examples to mitigate bias in existing classifiers with little reliance on existing ontologies?

To this end, first, we proposed CEnt, a novel entropy-based method that sup-

ports users with a set of actionable input alternatives to improve their outcomes. CEnt operates in a model-agnostic fashion while respecting immutability and semi-immutability constraints and encouraging attainable and plausible contrasts through a manifold-like distance metric. Our method improves the proximity and attainability of contrastive explanations without compromising latency and constraint violation. We utilized the derived contrasts to faithfully evaluate the fairness of a classifier, with our FCF metric, by first defining faithfulness guarantees for fairness and exploiting VAE distances to derive attainable contrasts. FCF is faithful to the individuals by contrasting them to neighbors that are less likely to be outliers and unattainable.

Then, we extended our contrastive fairness evaluation to textual settings by suggesting two novel deep inspection techniques for protected attributes and extending faithfulness guarantees to NLP. Our method, CoFE, is shown to expose discrimination that is diluted, even missed, by other fairness metrics on a variety of protected attributes. Furthermore, we defined faithfulness guarantees in textual settings. Our metrics enable us to advise on the use of latent representations in the generation of textual contrastive examples to improve faithfulness by 33%.

Finally, we complemented our study with a novel latent mitigation technique for textual classifiers. Our method does not rely on an extensive manually labeled ontology of analogies between privileged and underprivileged groups. We empirically validate its effectiveness in neutralizing transformers with respect to different sensitive attributes. Quantitatively, CoFE augmentation reduces bias in textual classifiers by more than 50% on average compared to 20% or 30% with existing methods.

The findings of this dissertation motivate the use of manifold-like distance metrics in the derivation of contrasts in the explainable AI field and in fairness evaluation. By utilizing VAEs contrastive examples are more likely to guarantee attainability

and plausibility ensuring a faithful counterfactual explanation and individual fairness evaluation. We also lay the foundation for a new era of classifier neutralization that does not heavily rely on existing ontologies and can be easily extended to new languages and sensitive attributes.

We acknowledge the limitation of any fairness metric, hence FCF and CoFE, in capturing all notions of bias. Our method, solely, evaluates the decision-making process of a classifier from an individual fairness lens. It can thus put the practitioner's fingers on particular discrimination to circumvent the issue with our proposed latent mitigation.

A promising vein of research focuses on enhancing privacy guarantees of our method within explainable and fair AI. Additionally, an investigation of the applicability of our method on imagery datasets can further enhance the comprehensiveness of our work. Within textual settings, filtering based on scores such as BLEU and Self-BERT scores would further improve the plausibility of our derived counterfactuals and hence enhance our latent augmentation technique.

Appendix

Objective II

Synthetic Dataset Generation

We describe the data generation scheme for the synthetic experiment of Section 5.4.1.

We assumed different data distributions for the privileged and underprivileged groups. The underprivileged group data is generated following this equation:

$$x_2 = 0.3(x_1 + 1)^2 - 1$$

The privileged group data follows:

$$x_2 = -0.3(x_1 - 1)^2 + 1 \tag{1}$$

Both distributions are overlapped with random normal in the range $[-1, 1]$.

```
import numpy as np

N_samples = 600
np.random.seed(47)
x_underpriv = np.linspace(-3, 1, N_samples)
x_priv = np.linspace(-1,3,N_samples)
```

```
y_underpriv = 0.3*np.square(x_underpriv+1)-1
y_priv =-0.3*np.square(x_priv-1)+1

noise = np.random.normal(-0.1, 0.1, y_underpriv.shape)

y_underpriv = y_underpriv + noise
y_priv = y_priv + noise
```

The *make_moons* function of *sklearn* could have been useful in this experiment. However, we opt for our data generation scheme to have control over the curvature of the generated parabolas.

Experimental Details of Section 5.4.2

The ANN model has two hidden layers with 50 and 25 neurons respectively activated via *relu* and trained with Adam optimizer and a 10^{-5} learning rate. The LR model is trained through stochastic gradient descent on $l2$ cost. CARLA's [141] implementation of the counterfactual explainability is used with the following setup.

CEM:

```
batch_size: 1
kappa: 0.1
init_learning_rate: 0.01
binary_search_steps: 9
max_iterations: 100
initial_const: 10
beta: 0.9
gamma: 0.0
```

CLUE:

```
train_vae: True
width: 10
depth: 3
latent_dim: 12
batch_size: 64
epochs: 1
lr: 0.001
early_stop: 10
```

DICE:

```
posthoc_sparsity_param: 0
```

GS:

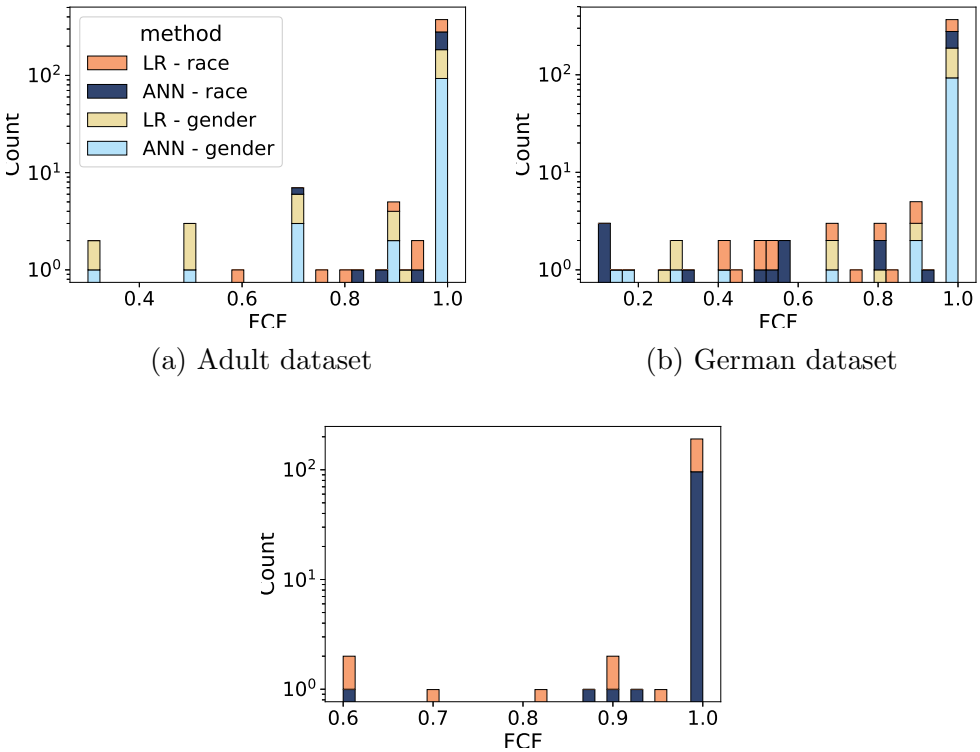
```
lambda: 0.5
optimizer: "adam"
lr: 0.1
max_iter: 1500
target_class: [0, 1]
binary_cat_features: True
```

Wachter:

```
loss_type: "BCE"
binary_cat_features: True
```

Distribution of Bias Scores

While we report the number of discrimination cases reported in each dataset in Section 5.4.4, we report the distribution of the continuous individual bias as computed by FCF in Fig. 1.



(a) Adult dataset

(b) German dataset

(c) COMPAS dataset

Figure 1: Distribution of FCF following log scale for the y-axis (higher scores indicate unbiased decisions)

Bibliography

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[2] M. Chen, A. Radford, R. Child, *et al.*, "Generative pretraining from pixels," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1691–1703.

[3] Y. Wu, M. Schuster, Z. Chen, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[4] A. Vaswani, S. Bengio, E. Brevdo, *et al.*, "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 2018, pp. 193–199.

[5] A. Fan, S. Bhosale, H. Schwenk, *et al.*, "Beyond english-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "Bert representations for video question answering," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1556–1565.

[8] S. Vakulenko, S. Longpre, Z. Tu, and R. Anantha, "Question rewriting for conversational question answering," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 355–363.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2157–2169.

[11] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[12] D. Silver, A. Huang, C. J. Maddison, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[13] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[14] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *International Conference on Machine Learning*, PMLR, 2019, pp. 803–811.

[15] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco, "Bias in word embeddings," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 446–457.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[17] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, 2018, pp. 0210–0215.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[21] M. Ribera and A. Lapedriza, "Can we do better explanations? a proposal of user-centered explainable ai.," in *IUI Workshops*, vol. 2327, 2019, p. 38.

[22] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.

[23] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[24] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[25] M. O'Shaughnessy, G. Canal, M. Connor, C. Rozell, and M. Davenport, "Generative causal explanations of black-box classifiers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5453–5467, 2020.

[26] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (cace)," *arXiv preprint arXiv:1907.07165*, 2019.

[27] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 607–617.

[28] A. Dhurandhar, P.-Y. Chen, R. Luss, *et al.*, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[29] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*, IEEE, 2018, pp. 1–7.

[30] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 656–666.

[31] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Information Processing Systems 30 (NIPS 2017) pre-proceedings*, vol. 30, 2017.

[32] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *Fed. Probation*, vol. 80, p. 38, 2016.

[33] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, eaao5580, 2018.

[34] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, vol. 67, 2017, p. 43.

[35] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 67–73.

[36] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang, "Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3320–3331.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[38] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[39] T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Issues with post-hoc counterfactual explanations: A discussion," in *ICML Workshop on Human in the Loop Learning (HILL 2019)*, 2019.

[40] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proceedings of The Web Conference 2020*, 2020, pp. 3126–3132.

[41] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[43] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[44] D. Zhang, F. Nan, X. Wei, *et al.*, "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*, 2021, pp. 5419–5430.

[45] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "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), Online: Association for Computational Linguistics, Aug. 2021, pp. 6707–6723. DOI: 10.18653/v1/2021.acl-long.523. [Online]. Available: https://aclanthology.org/2021.acl-long.523.

[46] G.-J. Qi and J. Luo, "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*, 2020.

[47] S. Sharma, J. Henderson, and J. Ghosh, "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*, 2020, pp. 166–172.

[48] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 219–226.

[49] P.-S. Huang, H. Zhang, R. Jiang, *et al.*, "Reducing sentiment bias in language models via counterfactual evaluation," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 65–83.

[50] M. Gardner, Y. Artzi, V. Basmov, *et al.*, "Evaluating models' local decision boundaries via contrast sets," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1307–1323. DOI: 10.18653/v1/2020.findings-emnlp.117. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.117.

[51] A. Artelt, F. Hinder, V. Vaquet, R. Feldhans, and B. Hammer, "Contrastive explanations for explaining model adaptations," in *International Work-Conference on Artificial Neural Networks*, Springer, 2021, pp. 101–112.

[52] P. Cheng, W. Hao, S. Yuan, S. Si, and L. Carin, "Fairfil: Contrastive neural debiasing method for pretrained text encoders," in *International Conference on Learning Representations*, 2020.

[53] P. Ma, S. Wang, and J. Liu, "Metamorphic testing and certified mitigation of fairness violations in nlp models," in *29th International Joint Conference on Artificial Intelligence*, 2020.

[54] N. Madaan, S. Mehta, T. Agrawaal, *et al.*, "Analyze, detect and remove gender stereotyping from bollywood movies," in *Conference on fairness, accountability and transparency*, PMLR, 2018, pp. 92–105.

[55] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[56] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[57] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.

[58] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[59] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri, "U-cam: Visual explanation using uncertainty based class activation maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7444–7453.

[60] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[61] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, 2017, pp. 6967–6976.

[62] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 233–255, 2016.

[63] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[64] K. Ethayarajh, "Rotate king to get queen: Word relationships as orthogonal transformations in embedding space," 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)*, 2019, pp. 3494–3499.

[65] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys (CSUR)*, 2022.

[66] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[67] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4593–4601.

[68] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 276–286.

[69] A. Raganato and J. Tiedemann, "An analysis of encoder representations in transformer-based machine translation," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 287–297.

[70] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[71] J. Lee, J.-H. Shin, and J.-S. Kim, "Interactive visualization and manipulation of attention-based neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2017, pp. 121–126.

[72] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "Seq2seq-vis: A visual debugging tool for sequence-to-sequence models," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 353–363, 2018.

[73] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 63–76.

[74] J. Vig, "Visualizing attention in transformer-based language representation models," *arXiv preprint arXiv:1904.02679*, 2019.

[75] A. Dhurandhar, P.-Y. Chen, R. Luss, *et al.*, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 590–601.

[76] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 344–350.

[77] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[78] T. Miller, "Contrastive explanation: A structural-model approach," *arXiv preprint arXiv:1811.03163*, 2018.

[79] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[80] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx, "Contrastive explanations with local foil trees," *arXiv preprint arXiv:1806.07470*, 2018.

[81] D. Dua and C. Graff, *UCI machine learning repository*, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[82] J. van der Waa, J. van Diggelen, K. v. d. Bosch, and M. Neerincx, "Contrastive explanations for reinforcement learning in terms of expected consequences," *arXiv preprint arXiv:1807.08706*, 2018.

[83] S. Rathi, "Generating counterfactual and contrastive explanations using shap," *arXiv preprint arXiv:1906.09293*, 2019.

[84] A. Dhurandhar, T. Pedapati, A. Balakrishnan, P.-Y. Chen, K. Shanmugam, and R. Puri, "Model agnostic contrastive explanations for structured data," *arXiv preprint arXiv:1906.00117*, 2019.

[85] M. T. Ribeiro, S. Singh, and C. Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, 2016. arXiv: 1602.04938 [cs.LG].

[86] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Inverse classification for comparison-based interpretability in machine learning," *stat*, vol. 1050, p. 22, 2017.

[87] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[88] M. Downs, J. L. Chu, Y. Yacoby, F. Doshi-Velez, and W. Pan, "Cruds: Counterfactual recourse using disentangled subspaces," *ICML WHI*, vol. 2020, pp. 1–23, 2020.

[89] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.

[90] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 895–905.

[91] M. T. Lash, Q. Lin, N. Street, J. G. Robinson, and J. Ohlmann, "Generalized inverse classification," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, SIAM, 2017, pp. 162–170.

[92] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Inverse classification for comparison-based interpretability in machine learning," *arXiv preprint arXiv:1712.08443*, 2017.

[93] R. M. Grath, L. Costabello, C. L. Van, *et al.*, "Interpretable credit application predictions with counterfactual explanations," *arXiv preprint arXiv:1811.05245*, 2018.

[94] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 903–912.

[95] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *CoRR*, vol. abs/1810.08810, 2018. arXiv: 1810.08810. [Online]. Available: http://arxiv.org/abs/1810.08810.

[96] Q. Hu and H. Rangwala, "Metric-free individual fairness with cooperative contextual bandits," in *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2020, pp. 182–191.

[97] Y. Zhou, P. Zhao, W. Tong, and Y. Zhu, "Cdl-gan: Contrastive distance learning generative adversarial network for image generation," *Applied Sciences*, vol. 11, no. 4, p. 1380, 2021.

[98] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," *Advances in neural information processing systems*, vol. 29, pp. 658–666, 2016.

[99] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, PMLR, 2016, pp. 1558–1566.

[100] M. Inácio, R. Izbicki, and B. Gyires-Tóth, "Distance assessment and analysis of high-dimensional samples using variational autoencoders," *Information Sciences*, vol. 557, pp. 407–420, 2021, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2020.06.065. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520306538.

[101] H. Ishfaq, A. Hoogi, and D. Rubin, "Tvae: Deep metric learning approach for variational autoencoder," in *Proc. ICLR Workshop*, 2018.

[102] S. Samanta, S. O'Hagan, N. Swainston, T. J. Roberts, and D. B. Kell, "Vaesim: A novel molecular similarity measure based on a variational autoencoder," *Molecules*, vol. 25, no. 15, p. 3446, 2020.

[103] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[104] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, pp. 4349–4357, 2016.

[105] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, Soviet Union, vol. 10, 1966, pp. 707–710.

[106] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, "Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning," in *ACL (1)*, 2016.

[107] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM journal on computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[108] Y. Zhu, S. Lu, L. Zheng, *et al.*, "Texygen: A benchmarking platform for text generation models," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1097–1100.

[109] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[110] D. Kahneman and A. Tversky, "The simulation heuristic.," Stanford Univ CA Dept of Psychology, Tech. Rep., 1981.

[111] A. Ross, A. Marasović, and M. E. Peters, "Explaining nlp models via minimal contrastive editing (mice)," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3840–3852.

[112] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "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*, 2020, pp. 119–126.

[113] R. Luss, P.-Y. Chen, A. Dhurandhar, *et al.*, "Leveraging latent features for local explanations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1139–1149.

[114] K. Natesan Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," *Advances in neural information processing systems*, vol. 33, pp. 5968–5979, 2020.

[115] C. Singh, W. J. Murdoch, and B. Yu, "Hierarchical interpretations for neural network predictions," in *International Conference on Learning Representations*, 2018.

[116] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Bia mitigation for machine learning classifiers: A comprehensive survey," *arXiv preprint arXiv:2207.07068*, 2022.

[117] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[118] H. Wang, B. Ustun, F. P. Calmon, and S. Harvard, "Avoiding disparate impact with counterfactual distributions," in *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*, 2018.

[119] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[120] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*, IEEE, 2009, pp. 13–18.

[121] L. E. Celis, V. Keswani, and N. Vishnoi, "Data preprocessing to mitigate bias: A maximum entropy based approach," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1349–1359.

[122] W. Du and X. Wu, "Fair and robust classification under sample selection bias," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2999–3003.

[123] D. McDuff, S. Ma, Y. Song, and A. Kapoor, "Characterizing bias in classifiers using generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[124] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency, "Towards debiasing sentence representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5502–5515.

[125] M. Kaneko and D. Bollegala, "Debiasing pre-trained contextualised embeddings," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1256–1266.

[126] S. Gillen, C. Jung, M. Kearns, and A. Roth, "Online learning with an unknown fairness metric," *Advances in neural information processing systems*, vol. 31, 2018.

[127] W. Zhang and J. C. Weiss, "Longitudinal fairness with censorship," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 12 235–12 243.

[128] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 2662–2670.

[129] Y. Roh, K. Lee, S. Whang, and C. Suh, "Fr-train: A mutual information-based approach to fair and robust training," in *International Conference on Machine Learning*, PMLR, 2020, pp. 8147–8157.

[130] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," in *International Conference on Learning Representations*, 2019.

[131] P. Lahoti, A. Beutel, J. Chen, *et al.*, "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[132] A. Petrović, M. Nikolić, S. Radovanović, B. Delibašić, and M. Jovanović, "Fair: Fair adversarial instance re-weighting," *Neurocomputing*, vol. 476, pp. 14–37, 2022.

[133] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[134] S. Chiappa, "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7801–7808.

[135] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2019, pp. 2847–2851.

[136] R. Mc Grath, L. Costabello, C. Le Van, *et al.*, "Interpretable credit application predictions with counterfactual explanations," in *NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*, 2018.

[137] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[138] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[139] R. Guo, P. Sun, E. Lindgren, *et al.*, "Accelerating large-scale inference with anisotropic vector quantization," in *International Conference on Machine Learning*, 2020. [Online]. Available: https://arxiv.org/abs/1908.10396.

[140] E. W. Dijkstra, "Oral history interview with edsger w. dijkstra," 2001.

[141] M. Pawelczyk, S. Bielawski, J. Van den Heuvel, T. Richter, and G. Kasneci, "Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms," 2021.

[142] W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 577–586.

[143] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a clue: A method for explaining uncertainty estimates," in *International Conference on Learning Representations*, 2020.

[144] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[145] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[146] M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive explanations in neural networks," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 3289–3293.

[147] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg, "Contrastive explanations for model interpretability," *arXiv preprint arXiv:2103.01378*, 2021.

[148] L. Faber, A. K. Moghaddam, and R. Wattenhofer, "Contrastive graph neural network explanation," *arXiv preprint arXiv:2010.13663*, 2020.

[149] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[150] P. Bobko and P. L. Roth, "The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice," in *Research in personnel and human resources management*, Emerald Group Publishing Limited, 2004.

[151] J. El Zini and M. Awad, "Beyond model interpretability: On the faithfulness and adversarial robustness of contrastive textual explanations," in *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[152] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[153] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, vol. 96, 1996, pp. 226–231.

[154] A. Dhurandhar, P.-Y. Chen, R. Luss, *et al.*, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 590–601.

[155] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a clue: A method for explaining uncertainty estimates," *arXiv preprint arXiv:2006.06848*, 2020.

[156] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Inverse classification for comparison-based interpretability in machine learning," *arXiv preprint arXiv:1712.08443*, 2017.

[157] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 915–924.

[158] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco, "Bias in word embeddings," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 446–457, ISBN: 9781450369367. DOI: 10.1145/3351095.3372843. [Online]. Available: https://doi.org/10.1145/3351095.3372843.

[159] L. Rhue, "Racial influence on automated perceptions of emotions," *Available at SSRN 3281765*, 2018.

[160] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. [Online]. Available: https://aclanthology.org/2020.acl-main.647.

[161] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.

[162] V. Zhelezniak, A. Savkov, and N. Hammerla, "Estimating mutual information between dense word embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8361–8371.

[163] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal processing*, vol. 16, no. 3, pp. 233–248, 1989.

[164] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, "A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula," *Human brain mapping*, vol. 38, no. 3, pp. 1541–1573, 2017.

[165] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, 1995.

[166] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. suppl_2, S231–S240, 2002.

[167] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, *et al.*, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2018.

[168] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*, PMLR, 2019, pp. 5171–5180.

[169] F. Pérez-Cruz, "Estimation of information theoretic measures for continuous random variables," *Advances in neural information processing systems*, vol. 21, 2008.

[170] M. De-Arteaga, A. Romanov, H. Wallach, *et al.*, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 120–128.

[171] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[172] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[173] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *arXiv preprint arXiv:2004.09297*, 2020.

[174] R. Guo, P. Sun, E. Lindgren, *et al.*, "Accelerating large-scale inference with anisotropic vector quantization," in *International Conference on Machine Learning*, PMLR, 2020, pp. 3887–3896.

[175] S. Krishna, R. Gupta, A. Verma, J. Dhamala, Y. Pruksachatkun, and K.-W. Chang, "Measuring fairness of text classifiers via prediction sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5830–5842.

[176] D. Slack, A. Hilgard, H. Lakkaraju, and S. Singh, "Counterfactual explanations can be manipulated," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[177] X. Han, T. Baldwin, and T. Cohn, "Diverse adversaries for mitigating bias in training," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2760–2765.

[178] M. Ye, C. Gong, and Q. Liu, "Safer: A structure-free approach for certified robustness to adversarial word substitutions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3465–3475.