Enhancing Robustness through Domain Faithful Machine Learning

by

Ananth Balashankar

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

New York University

May, 2022

Prof. Lakshminarayanan Subramanian

Dr. Alex Beutel

Acknowledgements

This body of research would not have been possible without the continuous and dedicated support of my advisors, collaborators, and family.

Abstract

In this thesis, we outline the research on the design of Domain Faithful Deep Learning Systems, that translate expert-understandable domain knowledge and constraints to be faithfully incorporated into learning deep learning models. In high-stakes domains like health, socio-economic inference, and content moderation, a fundamental roadblock for developing deep learning systems is that machine learning models' predictions diverge from established causal domain knowledge when deployed in the real world and fail to faithfully incorporate domain-specific structure in counterfactual data distributions. To overcome these limitations, we developed domain faithful deep learning systems through methodological contributions in ML model design, constrained optimization, data augmentation, and feature selection, for real-world applications. Specifically, we developed ML systems for consequential socio-technical and natural language understanding tasks by collaborating with domain experts and addressing critical research questions such as "What data distributions do domain practitioners care about?", "How to faithfully convert domain knowledge into model constraints for better generalization?" and finally "How to evaluate whether the ML models we learn are grounded in the domain knowledge and in what ways do they deviate?". The causal-aware and robust prediction models developed have shown that relying on data alone can lead to incorporating spurious correlations, and low accuracy in data-sparse or counterfactual scenarios, and hence, incorporating domain-specific structure in all stages of the machine learning pipeline is necessary for building robust predictive models.

Contents

List of Figures

List of Tables

1 Introduction

In this thesis, we focus on the design of **Domain Faithful Deep Learning Systems**, that translate expert-understandable domain knowledge and constraints to be faithfully incorporated into learning deep learning models. In high-stakes domains like health, socio-economic inference and content moderation, a fundamental roadblock for developing deep learning systems is that machine learning models' predictions diverge from established causal domain knowledge when deployed in the real world and fail to faithfully incorporate domain specific structure in counterfactual data distributions. Prior work in this space have formulated this problem as that of model generalization [298], data and label distribution change [251], domain adaptation [131], or adversarial robustness [70]. By doing so, they argue about model under-specification in the infinite data regime and data representativeness [75] over data distributions that are not realistically observed. While improving robustness of machine learning models is the core objective of all these approaches, they still fail to meet the expectations of domain experts on how machine learning models should behave when deployed in the real world.

Currently, domains where machine learning is being applied can be broadly distinguished based on the amount of prevalent enforceable domain knowledge in that domain. For example, causal models [328] are robust and compact representations of domain knowledge which have implications of the conditional probabilities of the effect given the treatment and covariate distributions. Such an abstraction is common and well understood in industrial settings where the data generating procedure is well documented. Causal knowledge is often expressed in various

forms - graphical causal models, semantic causal roles in sentences, theoretical model parameters. For example, causality based question answering lies at the core of customer support tools like chatbots. Prior ML models fail to capture the directed nature of causality, for example rain causes traffic delay, and not vice versa. Hence, learning asymmetric causal embeddings faithful to causal graphs can improve accuracy. Causal knowledge is also useful in data sparse conditions where interventions are often infeasible. For example, the task of forecasting famine is critical for the mobilization of aid to millions of people, but hard to solve due to data scarcity in fragile and poorer countries. By building a news-based causal-aware forecasting framework that extracts *causal features* from 11.2 million news articles across 2 decades in 15 fragile countries, we can improve forecasting accuracy compared to state-of-the-art predictive models.

On the other hand, even in domains where causal models are not established, certain counterfactual behavior of the machine learning models are expected. For example, trustworthy ML models in health recommendations need to be robust to medical concepts over unseen patient data, while traditional ML models focus only on optimizing accuracy over the observed but limited test data. By incorporating trust through doctor-specified mapping rules between diagnoses and medications through data augmentation, we can improve accuracy of state-of-the-art end-to-end neural models. Automated detection of online toxic comments improves the quality of interaction in social media. However, the variations in the context of comments make it hard to protect specific demographic groups from disparate impact. By explicitly modeling such nuances through *counterfactual data augmentation*, we can bridge the gap and improve the accuracy of detecting toxicity by 6%.

To overcome these limitations, I have developed domain faithful deep learning systems that directly incorporate domain knowledge in various stages of the machine learning pipeline - model design, constrained optimization, data augmentation and feature selection. This has led to deployments of domain faithful ML systems for consequential socio-technical and natural language understanding tasks by collaborating with domain experts. Specifically, we address critical re-

search questions such as "What data distributions do domain practitioners care about?", "How to faithfully convert domain knowledge into model constraints for better generalization?" and finally "How to evaluate whether the ML models we learn are grounded in the domain knowledge and in what ways do they deviate?". In doing so, we enable ML to be used towards positive socio-economic development, by tackling real-world societal problems in computational social science and NLP, and simultaneously addressing the fundamental ML research questions underlying these problems. Throughout this thesis, we adopt a research philosophy that strongly emphasizes "end-to-end system design", where algorithmic contributions are evaluated and deployed in the real world with the aim to adopt them at scale. For instance, the causal-aware and robust prediction models developed in collaboration with the World Bank and Google, have shown that relying on data alone can lead to incorporating spurious correlations, and low accuracy in data sparse or counterfactual scenarios, and hence, domain-specific structure is necessary for building robust predictive models. Overall, the research in the thesis has been focused on applying *domain faithful deep learning* to build causally faithful and heterogeneously robust predictive models in the domains of socio-economic inference, causal-aware deep learning, health, and toxicity detection. Each of these domains pose unique challenges on how to incorporate structure and the diverse techniques required to execute them. Now, we present the outline of the 4 sections of the thesis:

Domain Faithful Causal Models: Question Answering tasks power technologies like chatbots for customer support in businesses. Recent advances in machine learning for processing natural language text have broadly relied on large neural language models like Transformers which capture the relationships between the word tokens in long sequences. The fine-tuning of these language models for multiple tasks have demonstrated state-of-the-art performance on benchmarks like GLUE. However, these fine-tuned models perform poorly on counterfactual sentences or inconsistently on downstream tasks which have specific structure like graphical causal models

or domain-specific theory. In the causal-QA dataset [5], questions of the form "What causes X?" are posed, where X can be a disease, phenomenon and a real-world event. Neural Network models have been modified to predict causal links, but lack the consistency required, i.e undirected paths in a graph are still considered causal, whereas causal graphs are strictly directional. On the other hand, traditional Information Retrieval (IR) techniques that mine such causal information from knowledge graphs are limited in their generalizability to new and related terms mentioned in questions, i.e "flood" and "deluge" may have similar causes, but if "deluge" is not in the graph, then we have no way of estimating its cause. To overcome the limitations of using either an end-to-end model or domain knowledge as-is in its limited scale, we provide a way to incorporate the constraints imposed by the domain-specific structure - causal graphs in this case into BERT-like transformer based models. We demonstrate that when proximity between the embeddings of two nodes is modeled using a pseudo-quasi-metric, we are able to capture the directedness of causal graphs. Specifically, we measure three properties of *faithfulness* namely the uniformity of the embeddings, the correlation between distances of any two random nodes in the graph, and link prediction accuracy. In each of these graph-specific indicators, by imposing a regularization loss which penalizes inconsistencies in how the embeddings satisfy these two properties over two large causal graphs with 800K nodes, we obtain a fine-tuned embedding that not only achieves causal faithfulness better, but also improves the area under the Precision-Recall curve over the *Yahoo! Answers causal-QA dataset by 21%*.

Domain Faithful Feature Extraction: In socio-economic inference, the motivation is to have a broader positive societal impact using data-driven machine learning tools. Many applications which relied purely on data have faced issues as they did not incorporate domain-specific causal structure. For example, in the Flu prediction model based on Google Search Trends, it was shown that the model deviates over-time as compared to a one that incorporates signals derived from the Center for Disease Control (CDC). In the problem of predicting food insecurity task,

we overcome the challenge of data sparsity in fragile states which are often encumbered with infrastructural and conflict-based issues that makes the task of data collection harder. As traditional indicators like rainfall, vegetation index, etc are often delayed, we aim to use the news streams published by reputed sources like BBC, Reuters, AP, etc. to automatically extract and construct causally grounded indicators. Our contributions extend beyond the methodologies and have implications on the ethical and operational trade-offs a domain practitioner needs to make in a socio-technical system. In the famine prediction task, by extracting causes from scientific literature using Semantic Frame Parsing and then constructing time-series indicators by expanding to tokens with low Word-Mover distances, we are able to *reduce the food insecurity forecasting errors by 32%*. Additionally, alignment of models to domain expertise provides an additional incentive to practitioners - counterfactual reasoning: Not all episodes of famine are the same, and our methodology allows us to model what is the implication of each of the causes in improving the prediction accuracy at a fine-grained level of districts in 15 of the most fragile countries in the world over two decades.

Domain Specific Concordance and Counterfactual Robustness: Recent advances in applying AI for healthcare have often relied purely on data, but fail categorically when patients with different characteristics than the ones present in training data are presented. Specifically, in the medication recommendation task, learning end-to-end neural models based on historical electronic health records might prove to be accurate, but may not inculcate trust in doctors, unless the ontologies of medicine that are used as standards by trusted medical associations are incorporated. In the medication recommendation task, since all possible diagnoses that may be relevant might not be present in the training data, we improve the neural network model - G-BERT's *domain-specific concordance* based on expert-specified medical ontologies like medication and diagnostic code hierarchies and the mapping rules between them. By incorporating causal structure into machine learning models through categorical counterfactual data augmentation

and regularization, we guard against predictions that violate the domain knowledge over categories and improve the *categorical robustness of prediction models by 1.2x* and accuracy by 12% on the MIMIC-III dataset, as we rely less on spurious correlations in the data.

Further, in the domain of toxicity detection in online social media comments, social-science experts have long advocated for incorporating how specific demographic groups are susceptible to specific types of toxic comments. It is important to model secondary attributes that are relevant to the toxicity of a sentence explicitly when we aim to be fair based on demographic groups. In this scenario, one needs to be aware of group-specific language, idioms, quirks, and background history to ascertain the toxicity of a comment. But this nuance was never captured explicitly in BERT-based neural network models. We incorporated this domain knowledge through counterfactual data augmentation that model secondary variables and were able to improve the ability to detect toxic comments for all demographic groups, specifically black women, who were susceptible to more directed toxic comments. By augmenting examples of directed toxicity in a weighted manner to demographic groups that are more exposed to such comments, we are able to classify toxicity better on all demographic groups. Without this nuance of how toxic comments vary, and just optimizing for overall absolute error, the toxicity detection model would disparately perform poorer on specific demographic groups unintentionally. Through intervention on secondary attributes through counterfactual data augmentation, we not only improved the model's understanding of what constitutes toxicity, but also improved the accuracy on all demographic groups by 7%. This application clearly demonstrates that as a text classification model is scaled to be applicable to all demographic groups in a society, the secondary effects of covariates and how they impact the performance of a ML system depends on domain knowledge, and needs carefully expert supervision. Such business decisions and design choices have the capacity to influence the product experience for billions of users.

Domain Faithful Evaluation: Domain practitioners have often minimal guidance on the choice of parameters that AI tools in healthcare operate over. For example, in the angiographic

disease status prediction task, the variability of diagnostic features in different demographic groups is well studied. Here, practitioners need to carefully evaluate the trade-offs between the per-group accuracy across demographic groups, when an end-to-end jointly trained model is used. When we analyze the performance of ML models on specific demographic groups, we outline the choice of parameters of fairness and accuracy trade-offs that practitioners have based on Pareto Efficiency. For example, how accurate an ML model should be over patients with darker skin tone than lighter skin tone in a heart disease status prediction model is a choice that cannot be made blindly, but with careful consideration of the medical diagnostic equipment's characteristics and the Pareto optimality of the model's performance across demographic groups. Through the principle of Pareto Efficiency, we can potentially *improve group-level accuracies by 9.6%* on UCI datasets. Acting blindly based on the neural model's decisions in high-stakes scenarios might be sub-optimal and using our methodology, experts can now justify their choice, in case they were to be contested.

Part I

Improving Robustness through Domain Faithful Causal Models

2 | Learning Causal Graph Faithful Language Representations

2.1 Introduction

Learning distributed word representations that capture causal relationships are useful for real-world natural language processing tasks [359, 414, 127, 126]. Approximating the notion of causality with a similarity-based distance metric using separate vector representations for cause and effect tokens has led to significant improvement in the performance of downstream tasks like Question Answering, but can be too restrictive to generalize over unobserved edges in larger causal graphs [378]. In downstream causal reasoning based tasks like dialog systems [308], explanation generation [160], question answering [378], it is important to align the models with the corresponding causal graph. However, words that have low cosine similarity capture various semantic similarities, like relatedness, synonyms, replaceability, or complementarity, but not directionality [167]. Hence, any symmetric distance in an embedding space cannot convey the directed causal semantics for a downstream task [286]. In this paper, we overcome these two shortcomings and propose to optimize for directed *faithfulness* [391] that word embeddings have to satisfy towards a causal graph.

Prior work on capturing sufficient information for causal inference tasks from embeddings aims to directly use them for average treatment effect estimation [414]. We are, however, inter-

ested in a complementary question: "Can we learn word embeddings based on a distance measure that maps the directed distance between nodes in a causal graph to that in the embedding space?". Unlike prior work, which aims to learn a causal aware embedding restricted to direct link prediction [167], we propose faithfulness constraints so that causal word embeddings aims to preserve the partial ordering over pairwise distances in the directed causal graph. In this paper, to achieve the goal of learning faithful word embeddings with a vocabulary of more than 100K tokens, we minimize faithfulness violations over pairwise samples of nodes in the causal graph. Through this constrained optimization, we learn an embedding that can be applied directly for causal inference tasks but also generalizes to emergent causal links. It has been shown that NLP models need to understand such causal links that persist in the real world for safe deployment [127, 294]. Embeddings that violate the faithfulness property, can lead to spurious correlations based on co-location in the embedding space. For example, in a Yahoo! causal question-answering task's example: "What causes nosebleed?": the answers were "dry air", "heavy dust", "damaged nasal cells" and "liver problems". If we were to only rely on an undirected association based embeddings, the causes "dry air" and "liver problems" might be nearby (with distance of 2), but would be appropriately placed far in a *directed* causality based embedding space. To capture such asymmetric properties, we aim to preserve alignment with the causal graph by mapping causal links to an asymmetric quasi-pseudo distance measure during training to capture directionality of the causal graph as per Figure 2.1. Since human validated causal graphs can be used directly to answer questions of the type "What causes X?", we demonstrate the utility of learning faithful representations by using our distance-based features to solve the Yahoo! causal question-answering (QA) task. A causal QA task, unlike a standard QA task, can directly benefit from incorporating a causal graph into word embeddings to answer anti-causal queries. Our key contributions are:

- We define a faithfulness property for word embeddings over a causal graph, that captures geometric properties of the causal graph, beyond the direct link prediction by ensuring global proximity preservation.

- We propose a methodology to learn faithful embeddings through violation minimization which improves neighborhood detection by 31.3%, uniformity by 42.6%, and distance correlation by 54.2% using a quasi-pseudo distance metric.
- The faithful BERT and RoBERTa-based embeddings we learn, when used as inputs to a causal QA task, increases the precision of the first ranked answer (P@1) over existing baselines by 10.2%.

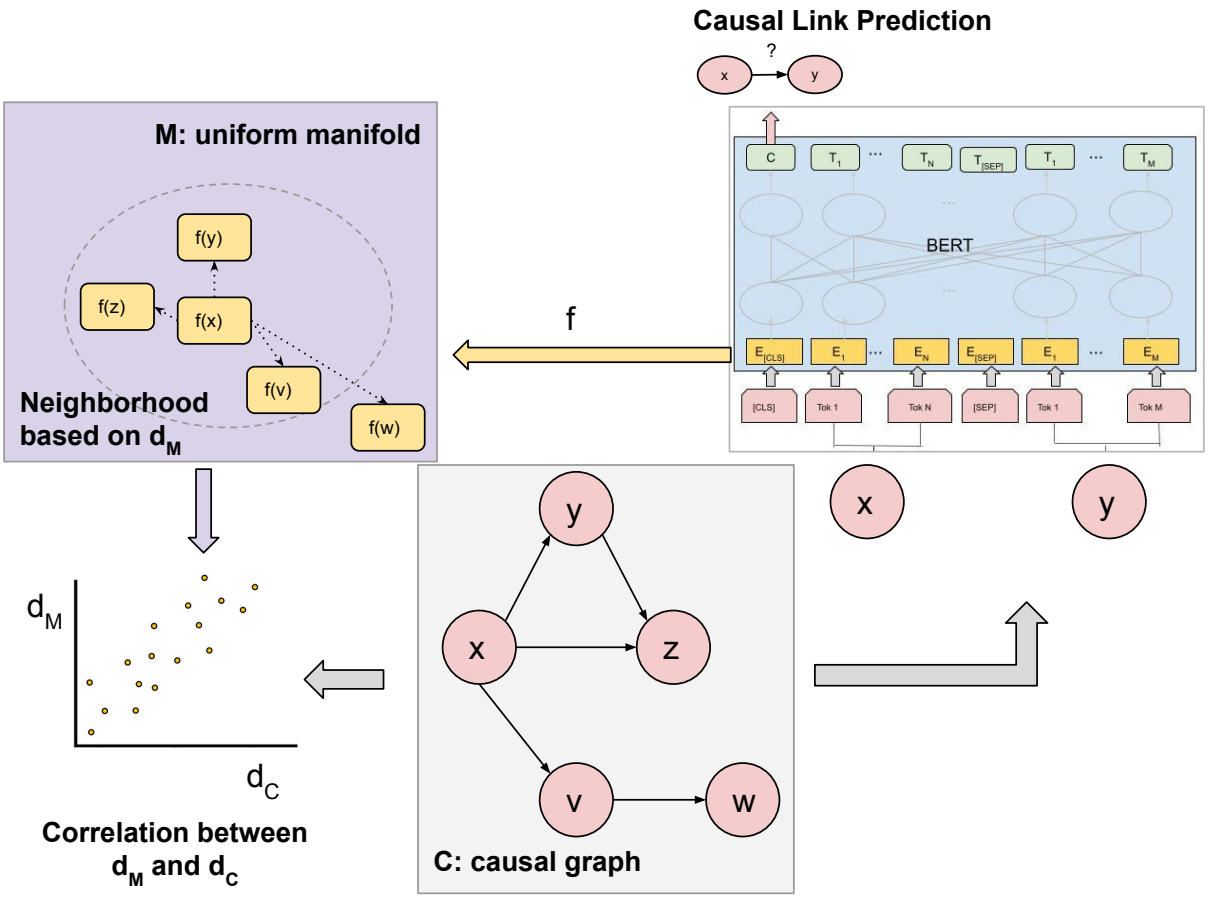


Figure 2.1: Schematic of our Faithful BERT-based model

2.2 Related Work

2.2.1 Causal Model Representations

Causal Inference, as outlined in [328] formalizes cause and effects discovered through intervention based experiments and communicates them via directed acyclic graphs. With the availability of large observational datasets for machine learning, various methods and assumptions have been proposed for learning causal graphs [372], data fusion and transportability properties [28, 48]. Specifically, our work closely aligns with the assumption of faithfulness [391], which requires that the observed probability distributions of nodes in a causal graph are conditionally independent as per the links in the graph. In our work, we use the probability distributions as modeled in a natural language model [231] and align it with the causal links in a graphical causal model. We extend the faithfulness assumption to be reflected in embeddings learnt by a masked language model [87, 255] for downstream tasks. This definition of faithfulness is different from the one proposed by [192] used to evaluate models for interpretability of models used for downstream tasks. Instead, our work builds on embeddings learnt in [378], given a causal model and learn embeddings that are boot-strapped using a small set of cause-effect seeds. Causal models have also been used to learn auxiliary tasks [110] using adversarial training to ensure that a language model learns causal-inspired representations. Such approaches use causal models to learn counterfactual embeddings invariant to the presence of confounding concepts in a sentence, while we encode the geometrical properties of causal graphs into the embeddings and the distance measure to maintain their faithfulness. In principle, we adopt a similar approach to [414] of fine-tuning towards a causal link prediction task. This is in contrast with approaches that use energy-based transition vectors used to represent the cause-to-effect and effect-to-cause links [467]. Our approach uses regularization constraints similar to the ones proposed for information bottlenecks in word embeddings [244, 153], text-based games [303], activation links in neuroscience [57], causal

consistency with ordinary differential equations [364] and temporal Granger Causality [400]. For an extensive survey of using text for causal inference tasks, we refer to [219].

2.2.2 Graph Representation Learning

Learning asymmetric transitive graph representations which generalize the causal graph have been studied extensively in Information Retrieval [60, 106, 247, 161]. They either utilize a random walk learning technique [331] or matrix factorization techniques [241, 402, 427, 291] to incorporate priors such as the stationary transition probability matrix, community structure, etc. More recently, [253, 321, 261] have incorporated knowledge graphs in BERT and shown increased accuracy in knowledge-centric NLP tasks. [470, 150, 322, 393, 398] propose asymmetric higher order proximity preserving graph embedding methods by learning separate source and target embeddings. While we can learn faithful 3-dimension embeddings for any fixed finite undirected graph deterministically [71], fine-tuning pre-trained word embeddings such that they generalize over all sub-graphs in a directed graph is known to be a hard graph kernel design problem that scales cubically with the number of nodes [420]. Our approach builds on efforts to incorporate graph-like structure in BERT, but overcomes the issue of learning dual embeddings for cause-effect edges by learning unified embeddings for both cause and effect roles of words. Through such embeddings, we can further aid causal discovery that is not yet captured in a graphical notation [61].

2.2.3 Graph Neural Networks

Recently, Graph neural networks that capture the graph neighborhood structure have been employed in link prediction [473, 2]. In [454], the problem is reduced to that of sequence prediction by reducing the graph to breadth-first search based deterministic sequence. In [246], node embeddings are updated after several rounds of message passing, while in [407] a variant of the random

walk is incorporated with a max-margin discriminative constraint. In [415], models are learned by attending over the neighborhood of nodes for context, while [223] apply spectral graph convolutions for a self-supervised learning task. We adopt the incremental approach proposed in [415] which does not rely on knowing the entire graph structure apriori and fine-tune on cause-effect pairs for the link prediction task on a pre-trained BERT-based language model.

2.3 Learning Faithful Embeddings

2.3.1 Background

Causal inference [328] aims to understand the cause and effect relationships between events. Learning purely based on correlations in observational data can lead to spurious causal links and can severely impact downstream tasks. Hence, intervention-based studies are conducted which carefully study the impact of a cause using controlled randomized experiments and other criterion to learn if links between causes and effects exist using observed data under specific assumptions. The findings of such studies are formalized using frameworks like Rubin Causal Models [365], Structural Causal Models [328], etc. While there are differences in abstractions between them, there is formal equivalence [124] in modeling counterfactuals ("What is the effect when the *cause* is intervened?") and we refer the reader to [330] for a primer in causal modeling.

In this paper, we assume a graphical structural causal model C [328] is given, whose nodes are linked with directed edges that denote the cause-effect relationship. For example, the cause-effect of "smoking" causes "cancer", references to the real world action of "smoking" in individuals that leads to the development of "cancer" kind of disease in those individuals. While causal models have a close relationship to the knowledge graph, the links of the causal graph have a well-defined causal interpretation that can be validated through counterfactual experiments. In this work, we assume the availability of such a causal graph and we do not aim to build one. Instead, we rely

on human annotators who with the help of web crawlers [177] and other information retrieval tools [378] produce a directed graphical causal model as shown in Figure 2.1.

2.3.2 Faithfulness

Given a graphical causal model C, we now present a faithfulness property an embedding that aims to closely align with the causal model has to satisfy. The faithfulness property was first proposed for any two causal spaces in [47] in the domain of quantum physics with the space-time dimension. Inspired by this, we propose an instantiation for word embeddings and a corresponding graphical causal model.

Definition 2.1 (Faithfulness)**.** An embedding $f : C \rightarrow M$ from a causal set (C, d_C) to a vector space (M, d_M) is faithful if:

- $\exists \lambda, \forall x, y \in C, d_C(x, y) = 1 \Leftrightarrow d_M(f(x), f(y)) \leq \lambda$
- $f(C)$ is distributed uniformly
- $\forall x, y, w, z \in C, d_C(x, y) \leq d_C(w, z) \Leftrightarrow d_M(f(x), f(y)) \leq d_M(f(w), f(z))$

Note that we use the causal set (C, d_C) as a tuple of the graphical causal model C and a distance measure d_C which is used to measure the *directed* distance between nodes in the graph. The vector space in which we map our embeddings is also characterized by a tuple (M, d_M), where M is the multidimensional real number space R^m, and a distance measure d_M which identifies nearby words in that vector space. The three conditions posed by the faithfulness property, more concretely specify that there needs to be a real threshold, within the embedding space, which can cover all the neighboring nodes of a word, the embedding space needs to be uniformly distributed, and finally, any inequality relationships between two distance measures in the causal graph needs to hold in the embedding space too. An embedding that satisfies this property can then be used to sufficiently represent the causal graph in downstream tasks.

2.3.2.1 Distance Measures

The definition of faithfulness is dependent on the distance measure used in both the causal graph and the embedding domains. In this work, we assume that the causal graph is a directed acyclic graph, and hence we measure d_C as the shortest directed distance (number of edges in an unweighted graph) between two nodes. If no such path exists between two nodes, we consider the distance to be a large number, which in the case of an unweighted graph, can be set to $> n$, where n is the number of nodes in the acyclic graph. Note that weighted graphs can also be incorporated with minor changes based on the maximum path in the graph.

However, the distance measure in the embedding space faces challenges in evaluation of simple supervised tasks [195]. To overcome these, we chose a distance measure that is closely tied to our faithfulness definition. We chose a unified set of embeddings for both the cause u and effect v, and, if there exists a causal edge from $u \rightarrow v$, then we would expect that $d_M(f(u), f(v)) << d_M(f(v), f(u))$. For this reason, symmetric distance choices like Euclidean distance, cosine similarity are not suitable. Our chosen distance measure, hence should follow the properties of quasi-pseudo metrics, defined as follows in [301]:

Definition 2.2 (Quasi-Pseudo Metric)**.** A measure $d_M : X \times X \rightarrow [0, \infty)$ is a quasi-pseudo metric if $\forall x, y, z \in X$,

- $d_M(x, y) \geq 0$
- $d_M(x, x) = 0$, but $d_M(x, y) = 0$ is possible for $x \neq y$
- $d_M(x, z) \leq d_M(x, y) + d_M(y, z)$

Hence, quasi-psuedo metrics, which do *not* satisfy the symmetry property are best suited to measure the distance between any two embeddings. We can generate such metrics, given a measure d. If the cause phrase u has p word tokens, and the effect phrase v has q word tokens,

we choose the Max-Matching method given in [447] in our definition of d_M by iterating through all pairs of words $(v_b, u_a) : v_b \neq u_a$. Note that the measure d computes the difference between v to u over the total m number of dimensions in $f(v_b), f(u_a)$.

$$d(u, v) = \min_{\substack{a=1..p \\ b=1..q \\ v_b \neq u_a}} \sum_{j=1}^{m} (f_j(v_b) - f_j(u_a)) \tag{2.1}$$

$$d_M(f(u), f(v)) = \begin{cases} d(u, v), & \text{if } d(u, v) > 0 \\ 10^{-d(u,v)} - 1, & \text{otherwise} \end{cases} \tag{2.2}$$

We chose this definition, as it is differentiable (except at 0, where we choose the gradient to be 0). Also, for each point u in the embedding space, there is a corresponding hyperplane that passes through it that defines the half-space which separates the reachable nodes $v : d(u, v) > 0$ - nodes which have either an indirect or direct causal link and the unreachable nodes $v : d(u, v) < 0$. Also, by the property of $d(u, v) = -d(v, u)$, we see that if v is reachable from u, then u is not reachable from v, thus affirming that this is suitable to represent a causal graph that is directed and acyclic.

2.3.3 Causal Graph Link Prediction

There are currently many approaches to learning causal representations, one which uses a masked language modeling approach where the word tokens in the cause are paired with word tokens in the effect using a skip-gram technique in an unsupervised setting. In the supervised setting, models align the cause-effect embeddings to solve either a sequence-to-sequence translation task or logistic classification task. Since we aim to capture all the nodes of the causal graph into a single set of word embeddings, we choose this approach. Further, in the supervised setting, we make explicit the causal relationship between cause and effect, thereby capturing the direction-

ality of the linkage. Thus, a supervised model could translate a cause to an effect or predict the link that exists from a cause to an effect. Among these supervised modeling choices, we choose the binary classification task of predicting if a directed edge exists between two nodes in the causal graph. This supervised learning is achieved by following the technique of fine-tuning as proposed in [414]. Formally, given a cause phrase u, an effect phrase v, let an $i(u, v)$ be an edge indicator variable $i(u, v) = \mathbb{1}_{u \to v}$ that takes binary values of $\{0, 1\}$ based on the existence of an edge from $u \to v$ in the causal graph.

Pre-trained Contextual Models: Pre-trained models based on transformers like BERT [87], RoBERTa [255] learn contextual embeddings of words or tokens by optimizing for the self-supervision task of predicting randomly masked tokens in a sentence. These pre-trained embeddings for word tokens have been used extensively for fine-tuning. Here, we use such fine-tuned models denoted as $\tilde{g}$ to predict the existence of an edge between the cause and effect u, v, by embedding them into $f(u)$, $f(v)$ respectively and further optimizing them in the fine-tuning stage on the following cross-entropy classification loss

$$\mathcal{L}_s = \mathbb{E}_{u,v \sim C} \, \text{CrossEnt}(i(u, v), \tilde{g}(u, v)) \tag{2.3}$$

2.3.4 Violation Minimization

Given the faithfulness definition, our goal is to learn an embedding that minimizes the number of violations of the faithfulness property. For each of the 3 conditions present in the faithfulness property, we define how we measure their adherence and incorporate it in the loss function. In addition to the causal graph link prediction task, we now present how the faithfulness properties are incorporated through regularization constraints.

2.3.4.1 Neighborhood

Since we expect a single embedding distance threshold that perfectly encapsulates the neighborhood of a node, we can measure this by varying distance thresholds for neighborhood detection and compute the area under the curve of the precision-recall curve. Since we aim to retain all the neighbors of a node in the causal graph within an upper bound of the distance in the embedding space, we add the sum of the distance between the nodes and their neighbors as an L1 regularization loss.

$$\mathcal{L}_n = \mathbb{E}_{\substack{u \sim C \\ v \in Neigh(u)}} |d_M(f(u), f(v))| \tag{2.4}$$

2.3.4.2 Uniformity

Since checking for true uniformity can be computationally intractable, we approximate by computing the per-dimension aggregate of all the word embeddings and compute the Wasserstein distance [312] between the observed distribution and the expected uniform distribution centered around zero (0^m). Since, in the uniformity constraint, we would expect that the embeddings are centered around zero, the mean of the embeddings should be close to zero. We measure the distance from this expected centroid and penalize the model for a high distance. If C_b denote the set of nodes chosen in a batch b, with size $|b|$, and $f_j(p)$ denote the j^{th} dimension of the embedding of node p, then we present the uniformity regularization loss:

$$\mathcal{L}_u = \sum_{j=1}^{m} \frac{1}{b} |\sum_{p \in C_b} f_j(p)| \tag{2.5}$$

2.3.4.3 Distance Correlation

To measure if inequalities between two distances in the causal graph hold in the embedding space, we measure the Pearson correlation coefficient between samples of distances between words in the causal graph and that of the embeddings. To ensure that any two distances sampled from the causal graph maintain the same inequality in the embedding space, we sample random nodes from the causal graph and compute the empirical Pearson Correlation Coefficient of their distances in the embedding space. A perfect correlation would lead to a coefficient of +1, so we penalize any deviation from that ideal correlation and present the distance correlation loss:

$$\begin{aligned} \mathcal{L}_c &= 1 - \rho_{d_C, d_M} \\ &= 1 - \frac{cov(d_C, d_M)}{\sigma_{d_C} \sigma_{d_M}} \end{aligned} \tag{2.5}$$

Note that all the above constraints are at a batch level and hence is added on to the batch cross-entropy loss during every back-propagation step. Since the losses are differentiable, we have used the auto-diff capability available in Tensorflow. The contribution of each of the above losses are combined using the Augmented Lagrangian method [182] and controlled using 3 parameters α, β, γ as follows:

$$\mathcal{L} = (1 - \alpha - \beta - \gamma)\mathcal{L}_s + \alpha\mathcal{L}_n + \beta\mathcal{L}_u + \gamma\mathcal{L}_c \tag{2.6}$$

The values of these hyperparameters were chosen to be $0.1, 0.15, 0.1$ respectively after cross-validation to optimize causal link prediction accuracy and faithfulness metrics. A summary of our approach is outlined in Algorithm 1.

The learning rate $a = 0.01$, $\mathcal{L}_u$, $\mathcal{L}_c$ are computed per batch by maintaining the required vari-

Algorithm 1 Faithful Embedding Training

Input: Pre-trained BERT based model $\tilde{g}$, causal graph C, distance measures: d_C, d_M,
for e=1..#*epochs* **do**
  $\mathcal{L} = 0$
  for j=1..b **do**
    $u, v \sim C : \sum \mathbb{1}_{i(u,v)=0} = \sum \mathbb{1}_{i(u,v)=1}$
    $\mathcal{L}_s$ += CrossEnt$(i(u,v), \tilde{g}(u,v))$
    $\mathcal{L}_n$ += $\sum_{w \in Neigh(u)} d_M(f(u), f(w))$
    Store $f(u), f(v)$ to update $\mathcal{L}_u$
    Store $d_C(u,v), d_M(f(u), f(v))$ to update $\mathcal{L}_c$
  end for
  Update $\mathcal{L}_u$, $\mathcal{L}_c$ and compute $\mathcal{L}$ (Eqn 2.6)
  Backprop $\tilde{g} \leftarrow \tilde{g} - a(\frac{\partial \mathcal{L}}{\partial \tilde{g}})$
end for

ables $f(u), f(v), d_C(u,v), d_M(f(u), f(v))$ in memory. These are implemented using Tensorflow's eager execution framework.

2.4 Evaluation

2.4.1 Causal Evidence Graphs

The causal evidence graphs we use contain phrases like "heavy rainfall" as causes and effects, which require us to learn the combined embeddings of the phrases. Restricting ourselves to just individual words would leave out the context required to understand the context to understand the cause-effect pairs. For example, the kind of effects "heavy rainfall" might have could be different from just "rainfall". We thus utilize the contextual embedding framework used to learn language models in BERT [87], as a way to learn contextual embeddings that align with a given graphical causal model. Note that there may be more than one causal model provided by experts based on their domains, and it is important to view our contribution as a way to align with domain expertise (for example, medical, legal, privacy, etc) with their respective causal models as a common mechanism to represent the said domain knowledge.

We use two causal graphs to construct their respective faithful embeddings, and demonstrate the utility of the embeddings in downstream tasks. The first causal graph we use is identical to the one used in [378], which uses the 815,233 cause-effect pairs extracted from the Annotated Gigaword and Wikipedia dataset, and an equal number of random relation pairs that are not causal as negative samples. The second causal graph is extracted from the web by [178], who use a bootstrapping approach with the initial pattern of "A causes B" and apply it to the ClueWeb12 web crawl dataset with 733,019,372 English web pages, between February and May 2012. From this web crawl, they provide a causal graph with 80,223 concept nodes and 199,803 causal links between the nodes. This graph has been sampled and validated by human annotators with over 96% precision. For our indirect evaluation based on downstream question answering tasks, we use the 3031 causal questions from Yahoo! Answers corpus [378]. These questions are of the form "What causes X?", and we use our faithful embeddings as a drop-in replacement for this causal QA task.

2.4.2 Metrics

Evaluating embeddings intrinsically has often led to varying leaderboards [195], hence we evaluate our embeddings based on their ability to map to the cause-effect relationship directly. We measure the faithfulness of the trained embeddings, using 3 metrics, one per property as per Eqns 2.4, 2.5, 2.5. For the neighborhood condition, we measure the area under the precision-recall curve as we choose multiple thresholds to define the neighborhood in the embedding space to correspondingly identify the relevant neighbors in the causal graph. For the uniformity condition, we measure the means of the per-dimension values of the word embeddings and compute the 1^{st} Wasserstein [312] distance from the expected centroid of zero. We also perform a statistical test for uniform distribution, which measures the mean Kolmogorov-Smirnov (K-S) test statistic [76] by bucketing embedding each dimension into 10 buckets. Since each dimension's test statistic can either pass or fail the test based on the significance level, we present the total number of

dimensions that pass the test at $\alpha = 0.05$ significance level. Finally, to measure the distance correlation property, we report the Pearson correlation coefficient between distances in the causal graph and the embeddings on a held-out part of the causal graph. For the QA task, we report the precision-at-one (P@1), the fraction of test samples where the highest ranked answer is relevant and the mean reciprocal rank (MRR) [270], the inverse of the position of the correct answer in our ranking on the held-out question set provided by [379].

2.4.3 Baselines

We evaluate our faithful embeddings by comparing them against two state-of-the-art approaches described in [378] and [414]. cEmbedBi uses a bi-directional model, with the task of predicting the masked cause and effect word tokens. This approach uses separate embeddings for words used as causes and effects. Causal-{BERT,RoBERTa} [414] uses the fine-tuning technique for the binary classification of edge detection, similar to ours, on the pre-trained large-uncased model. We can thus compare the gains we get by incorporating faithfulness conditions on the embeddings in downstream tasks.

2.5 Results

2.5.1 Faithfulness

As shown in Tables 2.1 and 2.2, our Faithful-RoBERTa model outperforms Causal-{BERT, RoBERTa} and cEmbedBi [378] on each of the three properties of faithfulness, namely the neighborhood, uniformity, and distance correlation, by more than 30%. Additionally, we report the correlation for Euclidean and Cosine similarity, despite not using it to optimize at training time. Faithful versions of the BERT and RoBERTa models increase the area under the curve of the precision-recall curve in detecting neighboring nodes of the Gigaword and CauseNet causal graphs by 21-23%

Embedding	Distance Correlation			Neighborhood
	Euclidean	Cosine	Quasi-Pseudo	AUC-PR
Gigaword Causal Graph				
cEmbedBi	0.33	0.48	0.52	0.67
Causal-BERT	0.40	0.55	0.61	0.71
Causal-RoBERTa	0.41	0.61	0.66	0.76
Faithful-BERT	0.42	0.63	0.78	0.88
Faithful-RoBERTa	**0.45**	**0.67**	**0.81**	**0.89**
CauseNet from ClueWeb12 web crawl				
cEmbedBi	0.23	0.37	0.34	0.54
Causal-BERT	0.25	0.38	0.39	0.56
Causal-RoBERTa	0.28	0.36	0.47	0.59
Faithful-BERT	0.31	0.41	0.55	0.68
Faithful-RoBERTa	**0.37**	**0.43**	**0.58**	**0.71**

Table 2.1: Correlation and Neighborhood faithfulness measures of the embeddings trained for both the Gigaword causal graph and ClueWeb12 CauseNet graph.

and 17-20% respectively. In Figure 2.2, we present the precision-recall curve when we use the models for ranking causal pairs above non-causal pairs on the SemEval Task 8 tuples [179] by varying the distance threshold in the embedding space which outlines the boundary of the neighboring nodes in the causal graph. This increase in accuracy for neighborhood detection indicates that incorporating the constraints during training time with our asymmetric causal embedding distance provides benefits in aligning the contextual embeddings as per the causal graph.

Embedding	1^{st}-Wasserstein	Mean K-S statistic	Uniform dimensions (1024)
cEmbedBi	0.54	0.54	205
Causal-BERT	0.45	0.43	348
Causal-RoBERTa	0.39	0.38	385
Faithful-BERT	0.31	0.21	541
Faithful-RoBERTa	**0.30**	**0.18**	**574**

Table 2.2: Uniformity measures on the embeddings learnt for Gigaword Causal Graph.

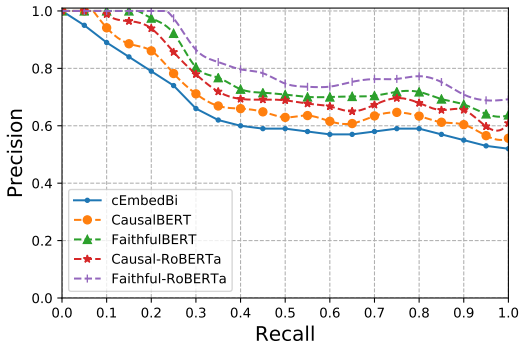


Figure 2.2: Precision-Recall to detect neighboring nodes in causal graph from the embeddings by applying threshold on distance measure

2.5.2 QA TASK

To evaluate if learning faithful embeddings is useful for causal aligned downstream tasks, we evaluate the fine-tuned embeddings to be directly used for question answering. As used in [120], we use the maximum, minimum, average distance between words of the question and answer words and the overall distance between the composite question and answer vectors from the embedding. Note that since both cEmbedBi and Causal-{BERT, RoBERTa} are trained with cosine similarity in mind, we use the cosine similarity, but for our Faithful-{BERT, RoBERTa} models, the distance measure used to rank is the quasi-pseudo metric defined in Def 2.2. We use these 4 features to train an SVM ranker to re-rank candidate answers provided by the candidate retrieval tool [194]. We see in Table 2.3 that Faithful-RoBERTa increases both the precision of the first answer predicted by 10.2%, and the mean reciprocal rank by 10.8%. This means that not only is the first ranked answer more causally correct, but the retrieval of the correct answer in the top-k positions has improved. This improvement in an out-of-domain QA task by aligning the embeddings to an externally available causal graph demonstrates that benefits of faithfulness transfer to downstream tasks.

Embedding	P@1	MRR
cEmbedBi	37.28	46.39
Causal-BERT	38.12	47.26
Causal-RoBERTa	38.74	49.01
Faithful-BERT	39.21	49.72
Faithful-RoBERTa	**41.07**	**51.42**
Ablation Study of Faithful-BERT		
w/o Neighborhood	38.55	48.67
w/o Uniformity	39.01	48.92
w/o Distance Correlation	38.28	48.04
Ablation Study of Faithful-RoBERTa		
w/o Neighborhood	39.69	49.39
w/o Uniformity	40.43	50.06
w/o Distance Correlation	39.50	49.28

Table 2.3: Performance on the QA task in Yahoo! Answers dataset using the Faithful versions of BERT and RoBERTa incorporating the Gigaword causal graph.

2.5.3 Re-alignment towards causation

To understand the reason behind the improved performance, we performed a qualitative inspection of 100 randomly sampled word pairs from the Gigaword causal graph [1] that are at varying distances in the original pre-trained embedding and trace how they have re-aligned after fine-tuning with the faithfulness objective. We annotate each of these word-pairs as being either causal or not as shown in the confusion matrix with examples in Table 2.4. In Figure 2.3, we see re-alignment of these word pairs from association based RoBERTa embeddings to the causally aligned Faithful-RoBERTa embedding space, that is, causal word pairs (blue and orange) move closer, and non-causal word pairs (green and red) move further based on the quasi-pseudo metric d_M. Specifically, the associative but non-causal word pairs (green) have moved further in Faithful-RoBERTa, while the non-associative but causal word pairs (orange) have moved closer. We see that in the cosine-similarity based RoBERTa, the causal word pairs had a mean distance of 0.48, while in the quasi-pseudo metric based Faithful-RoBERTa, the mean distance between the causal word pairs *reduced to 0.28*. The distances are normalized between 0 and 1 based on the maximum and minimum values of distances (cosine or d_M) in the sampled word-pairs.

We further analyzed how these associative and causal re-alignments impacted the causal QA task by categorizing the word pairs into three types of variables - mediators, colliders and con-

[1] https://github.com/ananthnyu/faithful-causal-rep/

	Cause	Non-cause
Associated	rain → flood	accident → fog
Non-Associated	war → epidemic	earthquake → spring

Table 2.4: Examples of word-pairs chosen to inspect faithfulness over the Gigaword causal graph.

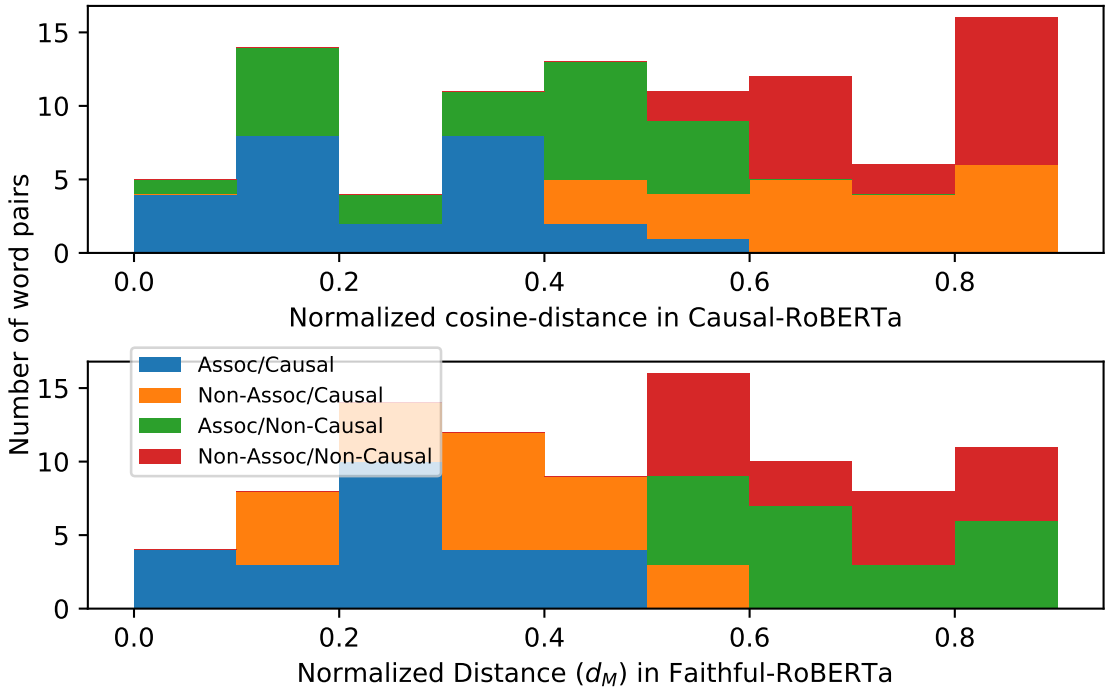


Figure 2.3: Re-alignment of word-pairs from the causal-RoBERTa embedding to our Faithful-RoBERTa (best viewed in color)

founders. **Mediators**: For the question, "What causes a tornado?", the answer involves "thunderstorms", which is a mediator caused by "high pressure". We see that "high pressure" is now much closer to "tornado" in Faithful-RoBERTa than baseline embeddings. **Colliders**: For the question, "What causes persistent cough?", the colliders "smoking" and "asthma" have moved further based on d_M in Faithful-RoBERTa. **Confounders**: For questions with confounders like, "What causes indigestion?", the confounding links "anxiety → indigestion", and "anxiety → insomnia" are near, but "insomnia → indigestion", is far. This further demonstrates the utility of incorporating faithfulness over multiple nodes of the graph, in addition to pairwise causal link

prediction.

2.6 Conclusion

We show that the faithfulness of text embeddings to a causal graph is important for causal inference-aligned downstream tasks. By incorporating the three faithfulness properties of neighborhood, uniformity, and distance correlation through regularization constraints while learning embeddings, we improve the precision of the first ranked answer in the causal QA task by 10.2%. We show that this is due to causal re-alignment of embeddings as per an asymmetric pseudo-distance metric.

Acknowledgments

We thank Sam Bowman for his feedback to the draft version of this manuscript. This work is partly supported by the funds from the Google Student Research Advisorship Program awarded to Ananth Balashankar.

3 Reconstructing the MERS disease outbreak using news

3.1 Introduction

The Middle Eastern Respiratory Syndrome - Corona Virus (MERS-CoV) disease is a new illness caused by a type of corona virus found in the Arabian peninsula since 2012. While most corona viruses have only cold-like symptoms, most people with the MERS virus had severe respiratory illness, gastrointestinal problems, sometimes leading to death [190]. As of end of September 2018, there were a total of 2260 laboratory confirmed cases and 803 associated deaths from MERS [320]. Despite the decreasing number of new cases over the years, WHO maintains its global risk assessment as it is mainly acquired from dromedary camels, a popular domesticated animal. There have been 218 instances of exported cases where contact with animals happened in the Middle East, but symptoms later manifest in the home countries of travellers. The difficulty in tracking MERS stems from the fact that, the dromedary camels show no symptoms when they are infected by MERS, making it harder to isolate them.

Early detection of MERS outbreaks is critical for health care resource allocation similar to diseases like malaria, dengue [1] and Ebola [74]. On the ground interventions can be mobilized in a more precise manner if the health agencies understand the local geographic, cultural and socio-economic conditions in a much fine-grained manner. However, structured signals on these

aspects are available yearly or quarterly through extensive surveys conducted by organizations like WHO and UNICEF [408], making it difficult to apply traditional machine learning techniques to predict outbreaks, which usually span a few weeks. For communicable diseases specifically, the mobility patterns of people and animals play an important role in determining the risk of an outbreak in a region and measuring this in regions with low access to tracking technology can be non-trivial.

In our study, we measure the factors that impact the propagation of the disease based on their mentions in the news. Specifically, we hypothesize that mobility patterns and access to local health care is impacted due to the presence of conflict within a region. This, in turn, influences the risk of a disease outbreak in a region. We use real time news streams such as GDELT [242] and the Uppsala Conflict data program [143] that aggregate statistics of conflict related death counts within a given geography. We use this localized knowledge in addition to a traditional disease transmission model for MERS [65] which estimates the susceptible, infected and recovered (SIR) number of people in a population based on the instrinsic characteristics of the disease as studied in a hospital. We extract interpretable variables, by running Granger Causality [155] tests for each of the hypothesized 56 news based indicators and keep only the ones which are statistically significant. We then embed the trained SIR model with the Granger-causal variables in a multivariate auto-regressive linear model to predict future infected number of cases and deaths.

Using sparse but rich conflict signals from the GDELT news database, our disease outbreak model is able to reconstruct the time series of actual infected cases as reported by WHO with a sum of squared errors which is 3.36x lower than using the standard MERS epidemiological model alone. The news based indicators which are most influential in our model represent the number of people killed, wounded and affected due to conflict in the regions of Lebanon, Kuwait, Egypt and Jordan. Some of these factors negatively influence the population mobility patterns and have disparate influence across regions. In addition to the variations of coefficients for news based

factors, we use sensitivity analysis and Granger Causal [155] time lags to interpret how each of these factors affect the timing and scale of the MERS outbreak in the middle east from 2013-2018.

3.2 Related Work

The environmental, animal and human transmission model [65] provided an understanding of how we could initialize the parameters for the transmission rate in the SIR Model. This work analyzed transmission patterns in a hospital in Saudi Arabia and identified the parameters of the SIR model. Apart from human-human transmissions, this model also incorporates animal-human interaction, especially from dromedary camels which serve as a large reservoir for the transmission of this disease. Incorporating this transmission alongside the human transmission rate significantly improves the accuracy of the model. The WHO currently educates people in the region to stop using animal products which could have come in contact with these camels when an outbreak is imminent.

The Dynamical Transmission Model [453] provided a corroboration to our parameter estimates. The sensitivity analysis provides an overview as to how the parameters would fluctuate on each iteration, which is in line with the modeling based on [65]. These analyses determined the changes that a parameter has on a model and the key drivers in a model(this happens to be the transmission rate b)

News based indicators have been used to predict man-made disasters and other natural events which are worthy of global attention previously in [309]. The tool developed was used to aid journalists in tracking events of consequence from Twitter streams [146]. In our work, we rely on established news sources and their aggregations. Parsing social media feeds would require sophisticated tools to filter false positives and would remain the focus of our future research direction.

Other auxiliary data like internet search history [**influenzagoog**] and the web [72] have been

used for disease surveillance, but the limitations of a fully unsupervised system without validation can cause spurious correlations as noted in [237]. A more purposeful and dedicated system built for disease tracking have also been deployed in real world systems as shown in [1, 107] rely on time series of structured data collected by specialists who were trained for this specific purpose. In this work, we try to take a combined approach [118, 455] by relying on aggregated news data which is not only easy to scale, but also validated by tools known to journalists and conflict trackers like the Uppsala conflict program. Thus, we aim to extract valid signals from a large news stream corpora to better understand disease transmission properties for MERS.

3.3 Background

In this section, we elaborate on the specifics of the MERS disease and motivate the need of news based modeling to overcome the challenges of addressing sparsity constraints in diseases like MERS. The hypothesis we will motivate in this section is that sparsity of on-the-ground signals relevant to disease modeling can be overcome by augmenting events from news which impact the migration and hence the disease propagation patterns indirectly. Specifically, we explore the scenario where conflict events impact the disease modeling of MERS in the Middle Eastern countries like Saudi Arabia, Kuwait, Lebanon, Egypt is presented here.

3.3.1 MERS

The Middle East Respiratory Syndrome is a respiratory illness caused by a coronavirus (MERS-CoV) and shows symptoms like fever, cough and shortness of breath. Close to 3-4 people who were infected have died of MERS related complications [190]. Although the disease was first reported in September 2012 in Saudi Arabia, it has since spread across the globe. In 2015, the largest outbreak outside the Arabian peninsula happened in South Korea and was traced back to a traveller from the middle east. MERS symptoms have been varied based on the risk factors

like diabetes, heart disease or weakened immune system. While severe complications including pneumonia or kidney failure have led to death, people who have shown milder symptoms or no symptoms have recovered. The incubation period of MERS is usually 5-6 days, but larger variations of 2-14 days have also been observed. This means that people who have come in contact with the virus can show no symptoms for up to 1-2 weeks [190]. This makes detecting MERS extremely difficult as it is known to have been transmitted through close contact with an infected person in addition to infected animals like dromedary camels, a popular animal for transportation in the middle east. Thus, 10 countries in the Arabian peninsula and 17 countries outside it have seen more than 2200 cases of MERS and there continues to be a threat of an outbreak.

3.3.2 Data Sparsity

As MERS is extremely hard to detect during the incubation period, many patients who show milder symptoms might go untested and can potentially infect people who have a higher risk of developing severe complications. Thus, the number of actual cases of MERS is harder to estimate due to lack of resources for testing and a lack of awareness. Thus, WHO and other health agencies rely on laboratory confirmed cases which form an extremely sparse data source. This will form the ground truth data in our analysis. Since the reports by the disease outbreak team by WHO are carefully cross-checked, it can be weeks or even months before the actual data is available for analysis. The reports are published weekly and sometimes fortnightly on the WHO's website [319] and is released widely. This limits the granularity of our analysis and rules out any real time analysis, daily or less based on streaming signals.

News signals about conflict are also considerably sparse with most coverage in the news relying on local sources that makes aggregation of data time-consuming. Press releases appear in batches, often with aggregate numbers over a longer time window. However, even such sparse news reports capture rich signals of conflict which can be specifically useful to predict the impact

on human migration. For example, initial death counts from a conflict gets reported on the day of the event, but the actual numbers are usually updated once more information is learnt and the corresponding statistics are updated. Relying on such sparse corrected sources is much useful than trying to parse all the available data, which can contain false information. We use such rich time series which are curated and verified by journalists and agencies on the ground for our analysis.

Given these sparsity constraints, we aim to reconstruct the time series of the actual number of infected MERS cases based on richer signals extracted from domain-specific knowledge about conflict and the corresponding limited data in the news.

3.3.3 Disease Outbreak Modeling

Traditional disease outbreak modeling relies on developing a mathematical model which denotes the rates of susceptibility (S), infection (I) and recovery (R) of a disease. This is usually modeled as differential equations where the assumptions are embedded in the way the equations are parameterized [65]. For example, for incurable diseases, recovery (R) is not modeled at all and sometimes, more than one type of infected and susceptible populations are tracked separately based on the mode of disease propagation. These assumptions stem from biological laboratory research which study the intrinsic propagation properties of a disease. Once such a mathematical epidemiological model is constructed by enumerating the number of compartments (S, I, R) and their interactions, its parameters are estimated and validated by a case study of a few specific hospitals and their surrounding regions. A critical parameter estimated through such case studies is the disease's basic reproduction number (R_o), which signifies the risk of the disease becoming an outbreak in a population [98]. An $R_o > 1$, indicates that unless sufficient interventions are not carried out, there would be an exponential increase in the infected population through a

multiplicative effect. Mathematically,

$$R_o = \rho(FV^{-1})$$

where ρ is the spectral radius of the next generation matrix, where F is a column matrix denoting the rate of increase in population of compartments and V is the rate of decrease in population from compartments due to all other causes.

3.3.4 Sensitivity Analysis

Once the parameters of the differential equations are estimated, a thorough sensitivity analysis of the parameters is done to understand how changing any of these parameters affects the susceptible, infected and recovered populations. Mathematically, this is done using the sensitivity index relative to the reproduction number for parameter p [369],

$$SI_{R_o,p} = \frac{\partial R_o}{\partial p} * \frac{p}{R_o}$$

Higher the $SI_{R_o,p}$, higher is the impact obtained by interventions that influence that parameter. One of the critical assumptions made while such models are used in practice is that the surrounding socio-economic, political and infrastructural environments of the place where the study was conducted and where it is deployed are identical for all matters concerning the spread of a disease. This inherently ignores the changes in the availability of health care and other such extrinsic factors. This deviation of the on-ground reality and conditions of case studies significantly impacts the efficacy of such models. Large data sets of these extrinsic signals are also not readily available in the regions which are most at risk of disease outbreaks.

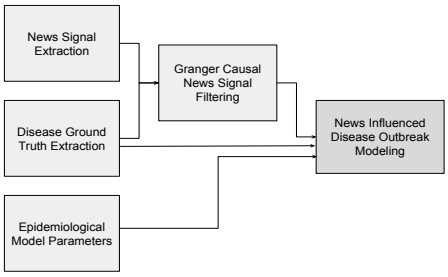


Figure 3.1: Outline of News-influenced Disease Outbreak Modeling

3.4 Methodology

In order to overcome this compounded problem of not being able to scale the epidemiology model to regions which are most at risk, due to the lack of extrinsic knowledge of socio-economic conditions in those fragile states, we resort to the news to extract meaningful signals for disease modeling. However, not all event-indicators in the news are relevant to disease modeling and careful inspection of the variables chosen is required. Hence, we take a conservative approach and filter only those variables related to the factors studied by social researchers for disease outbreak modeling and prescribed by WHO [317]. As per WHO, conflict is the primary factor that increases the risk of spread of infectious diseases like MERS. Hence, early indicators of even such sparse conflict related signals from news streams can significantly boost the accuracy of the SIR model applied for infectious diseases. In the remaining sections, we describe the methodology of our news based models and results.

Building news based models for disease outbreak modeling requires information retrieval tools to extract signals from the news, ground truth data from trusted sources, domain knowledge of the disease captured in graphical models of disease propagation and finally the prediction model which integrates all of this to produce the final estimate of the number of people infected by the disease. This is illustrated in Figure 3.1.

3.4.1 News Extraction

In order to incorporate news based signals in our disease outbreak prediction modeling to model extrinsic factors, we need to convert the words present in the news to a suitable representation which can capture the trends in the news. We hence chose to model it as a time series of events and its relevant statistics which are relevant to the disease. For example, for a regional *conflict* which is causing stress, we take into account the number of times that conflict was mentioned in the news and its associated number of deaths, wounded and sickened people. The definition of conflict can be ambiguous depending on the stakeholders and this extraction is conditioned on the domain expertise of the journalists in ensuring that aggregate statistics are not duplicated. These are usually extracted from the news article where it was mentioned. Quite often, the statistics reported are cumulative instead of the incremental change required at time t and hence we needed to build suitable tools and language filters to prune them.

In addition to raw news articles, we also used structured tables which are curated by organizations like Uppsala Conflict Program [333] to extract some of these relevant news signals. These are again suitably filtered using data processing tools. Once these time series were generated, they were normalized such that the time series is centered. This is required so that the variations in raw values across regions are comparable and are not dominated by the largest value. Any time series prediction task does not usually converge unless the time series is stationary and lack seasonal trends. To remove such trends, it is common to take differences until the final time series is stationary. However, in the case of sparse time series where conflict occurs based on seasonal and other trends which are not stationary, we resort to time series chunking. Each time series chunk, denoted by a start and an end date corresponds to a conflict episode and the time series within each episode is ensured to be stationary. We use such time series chunks throughout our prediction task.

3.4.2 Disease Ground Truth Extraction

Extracting ground truth of the number of cases and deaths associated to a disease can be quite controversial due to differing reports in the news and medical agencies. We rely on trusted sources like the UN, WHO to provide us with these estimates based on on-the-ground healthcare personnel. Some of these trusted sources provide data in the form of monthly reports or bulletins in the form of natural language text. We parse this text and extract relevant statistics like time, number of new cases and deaths reported for a disease across regions. Extraction is done using regular expressions as most of this text usually follows a template, which can be easily reverse-engineered. This provides the time series of the ground truth for the prediction task.

In order to take into account data outages and changes in template, we utilized RSS feeds on the disease outbreak portals to cross check the numbers extracted. These usually serve as efficient notifications of updates, but need to be monitored for changes undetected by web scrapers. Scaling these scrapers to multiple sources and in multiple languages remains out of scope for this task. However, while inspecting the news articles cited in these trusted disease outbreak sources, we usually noted that they were in the local language. Incorporating signals from these would be immensely useful for early-detection of outbreaks.

3.4.3 Epidemiological Modeling

Disease modeling based on rates of changes in population sizes at different stages of a disease is a common mathematical modeling approach. In this model, populations are compartmentalized and the rate of transfer of individuals from one compartment to another is modeled using differential equations. This can be easily visualized in a graphical model with each node denoting a compartment and the weights of the directed edges denoting the transfer rates. Each compartment is semantically annotated with a stage in their exposure to the disease like “susceptible”, i.e sub-population which is at risk of getting the infection, “infected” who have the infection

and "recovered" who have either recovered or died from the infection. In all such compartmentalized models, the population of the region is assumed to be constant and transitions between compartments have Markov assumptions.

This makes it easier to denote the graphical model in terms of differential equations with the rates of transfer and the nodes in the model being specific to a disease. MERS being an infectious disease has been studied by epidemiologists and several models have been proposed including SISI and SIR models [453]. SISI model, for example has two types of infections (primary and secondary) in two regions, where primary infections occur from contact with animals and secondary infections occur from contact with other infected humans in hospitals. The corresponding susceptible (S) and infected (I) populations are estimated using links from $S \longrightarrow I \longrightarrow S \longrightarrow I$.

In our modeling, we refer to the SIR (Susceptible, Infected, Recovered) model, a standard mathematical model which predicts how a disease propagates in a closed population over time. It represents the SIR population numbers as a function of time, and describes the time line of an epidemic, by fitting data from case studies on a small number of hospitals in the region where the disease is endemic. The sensitivity of this model is defined by the reproductive number (R_o) and the effect of MERS specific parameters on it are validated by epidemiologists on the population of Saudi Arabia [453]. We can relate the population numbers $s(t)$, $i(t)$ and $r(t)$ by the following differential equations. Solving for $i(t)$, given the initial population numbers, gives us the estimate of number of infected patients, which we refer to as $SIR[t]$ in the following sections.

$$\frac{\partial s}{\partial t} = -bs(t)i(t) \tag{3.1}$$

$$\frac{\partial i}{\partial t} = bs(t)i(t) - ki(t) \tag{3.2}$$

$$\frac{\partial r}{\partial t} = ki(t) \tag{3.3}$$

where b = rate of transmission, k = rate of recovery

3.4.4 Granger Causal Testing

Given two time-series X and Y , the Granger causality test checks whether the X is more effective in predicting Y than using just Y and if this holds then the test concludes X "Granger-causes" Y [155]. However, if both X and Y are driven by a common third process with different lags, one might still fail to reject the alternative hypothesis of Granger causality that X "does not Granger-cause" Y. Hence, in our modeling, we explore the possibility of causal links ignoring confounding variables due to the domain knowledge that there are no such confounding variable noted by the WHO. We note that if such an unobserved confounding variable exists, it is not considered in our Granger causality test.

In order to ensure that the news variables chosen are indeed related to the disease outbreak and not spurious correlations, we ran the Granger Causal test [155] between all of the news indicator variables (x) and the disease outbreak (y) as seen in Figures 3.2. We chose linear equations as our choice of modeling the prediction between x and y as it retains the benefits of interpretability in their coefficients. If, m, p, q denote the time lags of y, x in the auto-regressive equation at time t, then we can write:

$$y_t = a_0 + a_1 y_{t-1} + \ldots + a_m y_{t-m} + b_p x_{t-p} + \ldots + b_q x_{t-q} + error_t$$

Specifically, x is known to Granger-cause y, if there exists at least one non-zero coefficient of x which then leads to a significant improvement in prediction error over the case when we just use lagged values of y. We perform parametric F-tests on the non-zero coefficients of lagged variables and chose only the significant variables (p-values ≤ 0.05) to reject the null hypothesis that "the news indicator variable (x) does not Granger-cause the disease outbreak (y)". The chosen Granger Causal news variables are denoted by the vector $News[t]$ for a given time t, in the next sections.

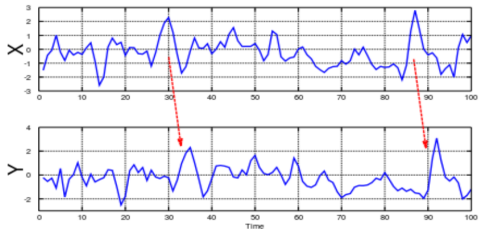


Figure 3.2: Granger causal link between two time series

Note that since true causality is hard to establish through observational studies, our goal here is to only find news variables which depict "predictive causality" and better predict future time series of the disease outbreak.

3.4.5 News Influenced SIR Modeling

Incorporating the news signals which are Granger Causal of the disease outbreak infections, into the epidemiological SIR model is the main methodological contribution of the paper. One option is to make changes to the equilibrium of the SIR model by altering the nodes in the graphical model and estimating the corresponding changes based on compartments induced by the news variables. This however does not scale to every disease specific model. Reconfiguring the disease model directly requires a lot of domain knowledge of both the disease and the related news variable, and remains out of scope of our paper.

Instead, we perceive the SIR model as yet another time series variable in a multivariate linear regression. This makes it possible to model other diseases easily in a similar manner without having to worry about the complex differential equations that govern the epidemiological transmission model of each disease. Now that we have the relevant news conflict variables chosen by the Granger Causal Test $News[t]$ and the MERS SIR model's value $SIR[t]$, we train a multivariate auto-regressive model with Lasso penalty [16] using glmnet [121] from lagged values of the ground truth $I_{t-\delta}$ and the regression variables as follows, where A, B, C are weight matrices,

maximum lag δ, and for any matrix x, let $x_{t-\delta} = x[t-\delta : t-1]$.

$$I[t] = A.I_{t-\delta} + B.News_{t-\delta} + C.SIR_{t-\delta} \tag{3.4}$$

$$\min_{A,B,C} \|I[t] - A.I_{t-\delta} - B.News_{t-\delta} - C.SIR_{t-\delta}\|_2^2 \tag{3.5}$$

$$\text{subject to} \|(A,B,C)\|_1 \leq r, \text{for a Lasso penalty } r. \tag{3.6}$$

The non-zero news coefficients that remain in the Lasso equation best explain the difference between SIR and ground truth in the News influenced disease model (Figure 3.5). The Lasso regularization embodies a variable selection procedure that ensures that only the most important variables are selected for prediction. We also reduce collinear variables in order to ensure that the Lasso regularizer does not pick variables which depict the same underlying event. This can be seen as a pre-processing step of removing a potential confounder variable as we cannot remove it once the regression model is trained. We use Variance Inflation Scores to prune out collinear variables [30]. This ensures that only those variables which cannot be estimated using the remaining news variables are used in the prediction task.

3.5 Evaluation

In this section, we explain the datasets used and the implementation details in the news influenced disease models.

3.5.1 Dataset

In this section, we describe the disease outbreak ground truth source and the news event databases used to extract conflict related signals in the region.

3.5.1.1 WHO-UN Dataset

The WHO-UN website [320] presents a collection of articles, which are updated every 8 to 15 days. Articles on each disease include statistics such as the number of cases or deaths and the date of the detected disease. The total size of this data spans 400 events for 192 countries from 2013 to 2018. There are 242 articles mentioning MERS, with breakdown of aggregate cases for each of the 12 Middle eastern regions + South Korea (traced back to a traveler from the Middle East). This data serves as our ground truth set.

3.5.1.2 GDELT Dataset

GDELT 1.0 Global Knowledge Graph [242] monitors the world's news from every country in over 100 languages with more than 1.5 billion events per year from April 2013 to Jan 2018, updated daily. These events are categorized based on killings or other crises such as natural disasters. It also provides a daily human count for each of these event types from sources like AFP, BBC monitoring, AP, WP, NYT and aggregator tools like Google News. We particularly focus on *killed, wounded, sickened and affected* events reported in each of the 12 regions as shown in Figure 3.3.

3.5.1.3 Uppsala Conflict Data Program

The Uppsala Conflict dataset [143, 333] provides deaths from organized violence keyed by a conflict ID and country, where each conflict has at least 25 related deaths in a year. The data set is presented as a time series with an yearly number of deaths per conflict. We focused on 8 of the 12 MERS regions which had a conflict (includes Saudi Arabia).

3.5.2 Data Preprocessing

We retrieved all the disease outbreak news articles from the UN website. These were later filtered to contain only the headline, timestamp, new cases and deaths using rule based string matching

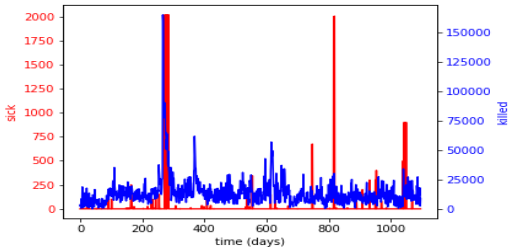


Figure 3.3: Killed and Sickened count in GDELT

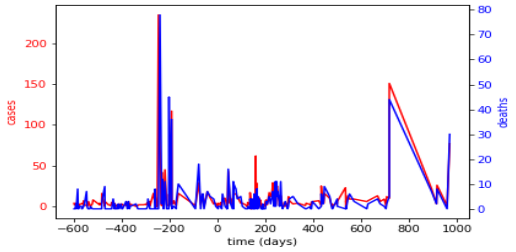


Figure 3.4: Number of new cases and deaths by MERS as reported by UN-WHO (> 70% in Saudi Arabia)

as can be seen in Figure 3.4. We extracted time series for each of the 48 normalized news indicator variables to range in $[-1, 1]$ for all (country, event-type) tuples from GDELT and 8 conflict variables per country from Uppsala. Time series chunking is also done to ensure that all the time series used for a specific time window is stationary. We take differences between consecutive values until stationarity is achieved. If we do not observe stationarity after differencing twice, we drop that time series from consideration as it no longer holds any interpretable meaning.

3.5.3 Model Parameters

The values of the SIR model's parameters as noted in Eqns [1-3] are predetermined. Specifically, the transmission rate $b = 1.4248$ and recovery rate $k = 0.1484$, used are based on the epidemiology study for MERS done in [65], as opposed to making theoretical estimations. The maximum time lag used for Granger Causal link estimation $\delta = 6$ weeks. The same maximum time lag was also used for the final news influenced multivariate auto-regressive model. This value was chosen based on the minimum size of the time chunk obtained in the data for 20 weeks.

3.5.4 Implementation

An overview of our implementation of building a news influenced disease model is given in Algorithm 2.

Algorithm 2 News Influenced Disease Modeling

Extract the ground truth timeseries for number of cases from the WHO-UN articles
Fit the epidemiological SIR model using pre-defined MERS specific parameters
Filter relevant conflict signals from GDELT and Uppsala by running Granger Causality tests
Train a multivariate auto-regressive model with SIR estimate and relevant conflict signals

3.6 Results

In this section, we will discuss the performance of the News Influenced disease model against several baselines. We pick 10 short outbreaks from 2013-2018, each spanning 21 weeks with the peak of the outbreak in the middle of the time series. The disease numbers reported are new cases and new deaths reported per week due of the disease. We fit the SIR model for each of these 10 outbreaks as per the variables mentioned above. We then normalize both the ground truth values and the SIR modeled values such that minimum and maximum values in the time series are scaled between 0 and 1 as seen in Figure 3.5. The final error calculated is the sum of the point-wise (one point per week) squared errors between the modeled and the ground truth. We report the average 10-fold cross validation error across multiple outbreaks.

3.6.1 Choice of News Source

In building a news based disease model, the source of the signals incorporated can have a significant impact on the trustability and accuracy of a model. Choosing between news sources can also influence the implementation requirements if this model were to be scaled. We tried vari-

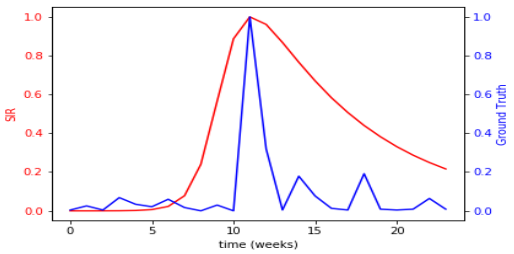


Figure 3.5: Difference between SIR and ground truth for an outbreak time window.

ous sources for the $News_{t-\delta}$ variable in Equation 3.4: 1) Conflict signals from GDELT 2) Conflict signals from Uppsala 3) Both GDELT and Uppsala 4) Only GDELT signals (no SIR, Uppsala). As mentioned in Table 3.1, SIR model with GDELT signals performs the best, reducing the error from the baseline SIR model of 8.99 to 2.68, an improvement by a factor of 3.36. The results presented in Table 3.1 are average errors from 10-fold cross validation of the episodes identified from time chunking. The low standard deviation of the errors shows that there is not a huge variation based on which chunks of outbreak episodes were used for training, indicating consistency and internal validity of the news influenced disease model.

Uppsala conflict signals were not useful for predicting the disease outbreak time series. We attribute this to the hand curated condensed extremely sparse (yearly) representation in the Uppsala event database. GDELT on the other hand is a daily aggregated database which captures the signals as represented in the news. This shows that GDELT has a better trade-off between aggregation coarseness and the time duration taken to put out verified conflict statistics. Another surprising result was that, using factors from GDELT alone in the multivariate auto-regressive prediction, produces a much lower error than the SIR model. This clearly indicates that local environmental and social factors are as important if not more important than the propagation properties of the disease within hospitals.

In Table 3.2, we see that , the news influenced SIR model performs well across outbreak episodes. The results presented are for those cross-validation rounds when the said episode was used for testing. The low variation is indicative that we can use the approach in predicting future

Model	Residual sum of squares error	Std deviation
SIR	8.99	0.65
SIR+GDELT	**2.68**	**0.42**
SIR+Uppsala	35.30	4.34
SIR+GDELT+Uppsala	2.81	0.43
GDELT	5.08	1.29

Table 3.1: Performance of News Influenced SIR Model

outbreaks and consistently explain the factors that were highlighted in the model.

Outbreak Episode	RMSE
March–June 2014	2.06
July–December 2014	0.40
January–April 2015	1.29
May–June 2015	2.38
June–July 2015	6.06
July–Sep 2015	1.92
April 2016 – August 2017	1.63

Table 3.2: Performance of News Influenced SIR Model across Episodes

3.6.2 Explainability of News Signals

Claiming lower prediction errors for the disease transmission patterns is not useful unless the model can be explained in terms of the multiple conflict signals in our model. Since, the time series used for analyzing each episode are normalized, we can directly compare the values of the coefficients. We chose the coefficients with the maximum absolute value over the many cross validation runs. This is highly correlated to the sensitivity index (SI_{R_o}) usually computed for disease outbreak models. The sign of the coefficients also indicate how conflict might indirectly influence the transmission patterns of the disease outbreak as can be seen in Table 3.3. In addition to the raw value of the coefficient, it is also useful to determine what is the expected time lag between a news signal appearing in the news and the expected influence on the number of infected people. This number (in weeks) when combined with the coefficient value, provides the estimate of when

and how much of an impact a signal in the news will have on the disease outbreak.

To illustrate this explainability, we choose to analyse the model predicting the outbreak from March-June 2014. Table 3.3 shows that although the time-lagged ground truth (actual counts from WHO) and SIR model remain the most important variables, conflict signals like kills in Kuwait and Lebanon (neighboring regions to Saudi Arabia) have a negative impact on the transmission of the disease, whereas increase in wounded and sick people in Egypt and affected people in Jordan indicate the increase in disease transmission of MERS. While events related to people being killed in conflicts could be traced to severe restriction of migration, while events related to being affected or wounded could seen as early indicators of people migrating due to the upcoming severe conflicts. While we note that there might be some feedback built into our model based on sick events retrospectively used, this requires further explorations.

Feature	Coefficient	Best time lag (weeks)
Lagged_Truth	0.17	1
SIR	0.23	3
kill_Kuwait	-0.17	5
kill_Lebanon	-0.15	5
wound_Egypt	0.12	5
affect_Jordan	0.10	1
sick_Egypt	0.03	1

Table 3.3: Important factors of the News Influenced SIR model

3.6.3 Implications

The above results which show more than 3x reduction in root means squared error is significant also because of the evidence it provides confirming the hypothesis articulated by WHO that conflict causes severe distress and exacerbates the spread of diseases. All the coefficients reported above are statistically significant (p-values < 0.05). Additionally, the time lags corresponding to each of the variables in multivariate regression provides us actionable information to facilitate timely interventions for disease containment. For example, when people in Jordan were affected

by conflict in March–June 2014, it led to an increase in MERS infected cases due to migration 1 week after the said conflict as illustrated in Table 3.3. Similar insights can be extracted for other outbreak episodes too. As per the current WHO fact sheet about MERS [320], there is no vaccine available for MERS, but appropriate hygiene needs to be practiced by people handling dromedary camels and the consumption of raw animal products should be minimized during the outbreak. Such advice is particularly useful for people affected in the region as symptoms of MERS appear later in the infection stage and is not easily distinguishable by health care workers. This early warning indicator is also beneficial for health care workers to prepare and use appropriate eye protection and other containment strategies including proactive blood tests.

3.6.4 News Sensitivity

Along with the timeliness of the news based disease model, we can also measure the sensitivity of the model for changes in the future related to conflict. This provides a way to distinguish the variations in the disease propagation pattern with any future significant escalation in conflict. We illustrate this sensitive analysis on the MERS outbreak from March–June 2014. Similar analyses can be done on other outbreak episodes too. We observe that even though some coefficients of news based variables are closer in value (kill_Kuwait and kill_Lebanon), the patterns they depict with respect to sensitivity significantly vary due to the underlying time series. For example, in Figure 3.6, we mostly see an increase in the number of MERS infected cases throughout the time series uniformly with increase in the number of people killed due to conflict in Kuwait. Whereas,in Figure 3.7 for Lebanon, we see both a phase shift and change in number of MERS infected cases with increase in number of people killed in conflict. We correspondingly see specific time periods where the impact is the highest from the conflict (week 9) as can be seen in Egypt for number of people wounded in conflict in Figure 3.8. Such variations in expected number of infected cases was not previously known or understood through time-based sensitivity analysis. This not only allows decision makers to categorize different types of conflicts, but also increases

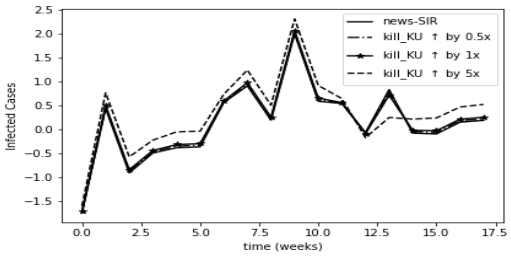


Figure 3.6: Sensitivity analysis based on number killed in conflicts in Kuwait indicates a uniform value shift in number of infected cases in March-June 2014.

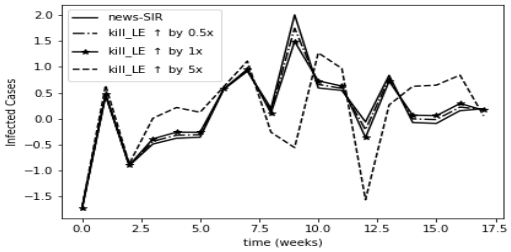


Figure 3.7: Sensitivity analysis based on number killed in conflicts in Lebanon indicates a disparate phase and value shift in number of infected cases in March-June 2014.

the awareness of the complexity and tight linkage between conflict and disease outbreaks.

3.7 Discussion

Is news a good modeling choice?: There is usually a disconnect between the disease modeling and the health care policy communities. While, the former relies on mathematical modeling to extract the most accurate parameters of the model, the latter cares more about adapting to on-the-

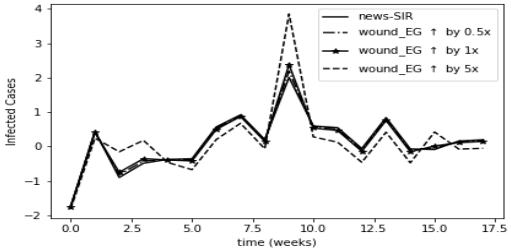


Figure 3.8: Sensitivity analysis based on number wounded in conflicts in Egypt shows varied shifts in number of infected cases at specific time intervals in March-June 2014.

ground realities and incorporating information on-the-fly into decision making. Mathematical models which are rigid and harder to interpret is usually not implemented by policy decision makers. This has led to customized web-tools built for practitioners to interface their knowledge with the underlying model [80]. We are inspired by such approaches and extend it to directly incorporate local information from the news. While news based modeling has the potential pitfall of relying on sentiments more than facts, we incorporate verified statistics about conflicts which get reported instead of the story around the event, which can be interpreted subjectively. This makes news based modeling a worthwhile choice for disease outbreaks which communicable through social contact.

Is MERS different than other diseases?: As MERS is heavily localized to countries in the Arabian Peninsula, it makes local news based modeling easier and drastically reduces the scope of news articles to be studied. MERS also has the clear distinction of a disease which spreads due to human and animal transportation in this region. This movement of people, animals and products is known to be a social indicator of the underlying political, economic and humanitarian conditions in the region. Thus, modeling MERS through news based modeling in the middle east makes more sense than other vector borne diseases or in any other region, outside the area of impact of the above macro-level events like conflict.

How can this be used at scale?: Having been able to reconstruct the time series of previous episodes of MERS with low average prediction error and low deviation in errors across all cross-validation of episodes, it provides us confidence to incorporate this model to predict future outbreaks. The model however might have to be tweaked to account for the efficient implementation of health care advisories issued by the WHO, which has significantly reduced the risk of MERS since it first occurred in 2012. This would impact the SIR component of the news influenced model, but not the factors learnt from the news, which are updated by design. Such a model, if adopted by the WHO or other health agency can significantly improve prediction of disease outbreaks based on historical patterns in the news, and lead to better intervention and

information dissemination strategies.

When would this model not work?: Further analysis however is required to breakdown the different types of conflict and the corresponding regions they impact. This can be done through spatial and text based categorization of the news articles which mention conflict. Such a model would however significantly suffer from the sparsity in the data post the categorization of conflict, similar to how news signals from the Uppsala Conflict Data program proved to be less effective due to the sparsity of data. This challenge needs further model improvements and cannot be addressed by the current news based model. One option we are actively pursuing is the tree-based factorization of the news signals which combines the best of both sub-categorization and larger datasets in a hierarchical approach.

3.8 Conclusion

Susceptibility, infection and recovery is modeled in disease transmission models using intrinsic properties of the disease. However, extrinsic factors also influence disease transmission and have been previously unexplored. We study the effect of regional conflict on the mobility patterns of people and animals for the transmission of MERS-CoV and show that by augmenting conflict based signals in real time news streams with a standard MERS SIR model, we significantly lower the infected population prediction error. Inspection of our news influenced disease model provides a human interpretable understanding even with very sparse signals.

Part II

Domain Faithful Feature Extraction

4 Extracting Causal factors from News Streams for Famine Forecasting

4.1 Introduction

Food insecurity continues to threaten the lives of hundreds of millions of people around the world today. According to the Food and Agriculture Organization of the United Nations (FAO), the number of undernourished increased from 624 million people in 2014 to 688 million in 2019 [109]. There is considerable evidence that quickly responding to emerging risks of food insecurity saves lives and lowers humanitarian costs [352], which leads aid agencies to resort to early warning systems to decide when and where to deploy emergency relief [23]. While risk factors are well-established [374, 285, 279, 287], ranging from conflicts to pests, weather shocks and migration, delayed or infrequent measurements of these factors typically impede early warning systems' ability to promptly anticipate food crises [14]. Furthermore, fragile states prone to food insecurity often lack the capacity to systematically measure risk factors, generating data gaps [441]. Against this backdrop, the past decade has seen an explosion in the availability of vast repositories of digital data, from satellite imagery to call detailed records, which are increasingly being analyzed to address socioeconomic challenges [196, 43]. Encouraged by these approaches,

we take advantage of recent advances in deep learning and natural language processing to extract anticipatory signals of food insecurity episodes from the text of a large corpus of news articles. Unlike existing food insecurity early warning systems, the articles we collect are published on a daily basis, allowing us to generate high-frequency forecasts [423]. News aggregators provide access to media articles curated in a transparent manner going back several decades, enabling the analysis of long time series of news streams [426]. Finally, authoritative media sources such as *BBC* or *Reuters* have a long-standing reputation for providing trustworthy information on local contexts, suited to produce disaggregated forecasts [38].

4.2 Background

Our study focuses on predicting the Integrated Phase Classification (IPC) of food insecurity. This classification is available at the district level across 37 fragile states in Africa, Asia, and Latin America, and was reported every four months between 2009 and 2015 and every three months thereafter. Food insecurity is classified according to an ordinal scale composed of five phases: minimal, stressed, crisis, emergency, and famine (supplementary text 4.3.2). We postulate that the factors triggering a food crisis are mentioned in the news prior to being potentially measured by traditional risk indicators. We therefore collect a novel dataset from the news aggregator Factiva containing 11.2 million articles covering countries included in the IPC dataset and published between June 1980 and July 2020 (supplementary text 4.3.3). We then develop a methodology based on semantic role labeling to elicit textual mentions of causes of food insecurity [24, 394]. We start from a handpicked list of 13 target keywords related to food insecurity and we use a frame-semantic parser to uncover causes of food insecurity appearing in the same semantic frame as one of the target keywords (supplementary text 4.4.1). For example, when the parser examines the sentence "*Famine may return to some parts of the country, with the eastern Pibor county, where floods and pests have ravaged crops, at particular risk*", it detects that "*floods*" and

"*pests*", both being established causes of food insecurity, are mentioned in the same semantic frame as the target keyword "*famine*". We apply this method on our corpus of news which allows us to elicit 1,062 text features consisting of unigrams, bigrams and trigrams occurring in the same semantic frame as a target keyword. To ensure that we capture a wide range of causes of food insecurity recognized both by journalists and by experts, we repeat the same procedure on a manually selected list of 93 peer-reviewed studies and books on food insecurity, which reveals 149 additional text features. We then expand this seed list of 1,211 features by considering text features mentioned in the news which are semantically similar to a seed [289, 234], obtaining 738 new features (supplementary text 4.4.2). Finally, to drop any irrelevant text feature that might have accidentally been picked up, we first convert each text feature into an index per month and per district by computing the proportion of monthly news articles mentioning both the text feature and the district ("news factor"). We then discard news factors which are not predictive of the IPC phase [222], leading to a final set of 167 text features (supplementary text 4.4.3). To shed light on these features, we partition them into 12 semantically distinct clusters (Fig. 4.10). We find that text features belonging to the same clusters tend to co-occur in the news, the average pairwise correlation between news factors in the same cluster being 69.9% versus 34.9% for those in different clusters, which provides support to our partitioning (Fig. 4.5). We also find that 9 of 12 clusters are composed of text features related to known causes of food insecurity – "conflict and violence", "political instability", "economic issues", "production shortage", "weather conditions", "land-related issues", "pests and diseases", "forced displacements", and "environmental issues" – accounting for 92% of the articles in which text features are mentioned. The remaining 3 clusters include terms related to "food crises", to "humanitarian aid", and "other" negative terms unspecific to food insecurity. Having established the consistency between our text features and known causes of food insecurity, we also demonstrate the presence of a strong cross-sectional relationship between news mentions and traditional measures of causes of food insecurity (Fig. 4.11). We use a comprehensive dataset containing traditional measures of food insecurity risk factors

("traditional factors") used in early-warning systems – a conflict fatality count, the change in food prices, an evapotranspiration index, a rainfall index, and an inverted vegetation index – covering 21 fragile states over the period 2009-2020 (supplementary text 4.5.1). After summarizing each traditional factor and each news factor at the district level with its maximum monthly value during the observation period, we associate with each traditional factor the news factor with which it has the highest Spearman correlation across districts (Fig. 4.11A-4.11J). We find that the conflict fatality count, change in food prices, evapotranspiration index, rainfall index, and inverted vegetation index are most strongly correlated to news mentions of "conflict", "food prices", "drought", "floods", and "pests" respectively ($r_S > 0.89$), thereby providing an additional sanity check for our approach (Fig. 4.11K-4.11O). Taken together, these results indicate that our procedure allows us to uncover text features that are consistent with established causes of food insecurity, interpretable, and validated by traditional risk indicators.

4.3 Dataset

4.3.1 Data collection

4.3.2 Food insecurity classification data

Our dataset on food insecurity comes from the Famine Early Warning System Network (FEWS NET). Food insecurity is classified into 5 phases following the Integrated Phase Classification (IPC) framework: (1) minimal, (2) stressed, (3) crisis, (4) emergency, and (5) famine. Phases are determined by experts and published at the district level across 37 countries since 2009, allowing us to compare food insecurity levels across time and regions in a standardized way. The classification covers the following countries: Afghanistan, Angola, Burundi, Central African Republic, Cameroon, Chad, Congo, El Salvador, Ethiopia, Guatemala, Guinea, Haiti, Honduras, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Nigeria, Niger, Rwanda, Senegal

Sierra Leone, Somalia, South Sudan, Sudan, Tajikistan, Djibouti, Tanzania, Uganda, Burkina Faso, Republic of Yemen, Democratic Republic of the Congo, Zambia, and Zimbabwe. It was established 4 times per year from 2009 to 2015 – in January, April, July, October – and 3 times per year thereafter – in February, June, October (Fig. 4.1).

4.3.3 Corpus of news articles

Our dataset of news articles comes from Factiva, a digital archive of global news content which aggregates more than 33,000 news resources from 200 countries in 28 languages. Each news article is tagged with geographic region codes, allowing us to ascertain its relevance to a specific country. We collect the text of the 11.2 million articles in English tagged with at least one of the 37 countries covered by FEWS NET (Fig. 4.2). While 60.5% of the articles were published by news sources located in a fragile state, the remaining 39.5% come from news sources based in a non-fragile state.

4.4 Feature selection

4.4.1 Frame-semantic parsing

We use a frame-semantic parser to extract from our corpus of news articles the text features which are causally related to food insecurity [24]. The parser first splits each sentence into syntactic constituents $c_1, c_2, ..., c_k$, where each c_i includes $p \geq 0$ contiguous word tokens $w_j, w_{j+1}, ..., w_{j+p}$ starting from position j. It then assigns to each syntactic constituent c_i a semantic role t_i. We use a deep neural network model reaching state-of-the-art accuracy on the benchmark dataset FrameNet to predict the semantic role of each syntactic constituent in our news corpus [394]. A semantic frame f is a collection of syntactic constituents along with their semantic roles $(c_i, t_i)_{i \in f}$. To select text features corresponding to causes of food insecurity, we restrict the set of semantic

frames produced by the parser using the following filters:

1. First, we exclude semantic frames whose constituents' roles do not include at least one "cause" and one "effect". Note that there might more than one "cause" and one "effect" per frame.

2. Next, we exclude semantic frames in which the "effect" constituents do not contain any of our 13 target keywords related to food insecurity (Fig. 4.3A).

3. Next, we exclude semantic frames which do not contain any of the causal links included in the FrameNet lexical database (Fig. 4.3B).

4. Finally, we select all the unigrams, bigrams, and trigrams mentioned either in a "cause" or in an "effect" constituent of the previously selected frames (Fig. 4.3C).

This procedure allows us to elicit 1,062 text features from our corpus of news articles. To ensure that our text features cover the causes of food insecurity established by experts, we also handpick a list of 93 well-cited books and peer-reviewed studies on food insecurity from Google Scholar using the following queries: "causes of famine", "food insecurity causes", "food insecurity Africa causes", "causes of food crisis", "famine Africa", "famine Africa causes" and "food crisis Africa causes". We then run the parser on the text of these books and studies which reveals 149 additional text features (Fig. 4.3D).

4.4.2 Keyword expansion

While our semantic parser allows us to extract a seed list of 1,211 features causally linked to food insecurity, it fails to capture words semantically close to a seed that are also relevant. For example, the parser selects the word "terrorism" but not the equally relevant word "terrorist" which does not appear in any of the causal frames that we considered. For this reason, we expand our seed list of features with words and phrases semantically close to each seed. We do so by considering

as candidate features all the unigrams in our news corpus as well as all the bigrams and trigrams occurring more than 1,000 times. We then convert the words of each feature into an embedding vector such that words occurring in similar contexts end up close to one another in the embedding space [289]. By computing the word mover's distance between each seed and each candidate feature, we keep the candidates whose distance to a seed is smaller than 6, obtaining 738 new features [234]. We find that expanding to more distant candidates does not lead to additional relevant features, which indicates that the features that we select cover a wide range of causes of food insecurity (Fig. 4.3C and Fig. 4.4).

4.4.3 Dimensionality reduction

Having uncovered a set of 1,949 text features causally related to food insecurity, we now aim to focus on those whose news mentions help predict the IPC phase. We first convert each text feature w into a time series $x_{w,d,t}$ measuring the proportion of news articles mentioning w and the district d during the month ending on date t ("news factor"). We then convert the IPC phase into a monthly indicator by forward filling the latest observation. We then estimate a panel autoregressive distributed lag model of the IPC phase $y_{d,t}$ in district d during the month ending on date t:

$$y_{d,t} = a_0 + a_1 y_{d,t-1} + a_2 y_{d,t-2} + \ldots + a_n y_{d,t-n} + b_1 x_{w,d,t-1} + \ldots + b_n x_{w,d,t-n}, \tag{4.1}$$

where the number of lags n is chosen based on the Akaike Information Criterion. This estimation is done at the district level which is the smallest geographic unit for which the IPC phases are observed. To determine whether a news factor predicts the IPC phase, we run a Granger-causality test and we reject the null hypothesis that x_w does not Granger-cause y if the news factor and its lagged values whose coefficients are statistically different from zero add explanatory power to

the regression according to an F-test at the 1% level. Since the Granger causality test assumes stationarity, we take the first difference of non-stationary time series until it passes the Augmented Dickey-Fuller test. If a news factor is predictive of the IPC phase after differentiation, we keep the transformed news factor for the rest of the analysis. Out of the 1,949 text features previously selected, we retain the 167 text features which are predictive of the IPC phase (Fig. 4.3C).

4.5 Predicting food insecurity

4.5.1 Traditional risk indicators

To uncover the predictive value of our news factors, we compare them with traditional measures of food insecurity risk factors obtained from recent studies on food insecurity [14, 423, 13]. Our risk indicators include:

- a monthly count of violent conflict events and the monthly average number of fatalities per event
- a food prices index (monthly log nominal food price index and monthly year-on-year difference)
- an evapotranspiration index (monthly mean)
- a rainfall index (monthly mean and deviation from average seasonal value)
- a normalized difference vegetation index (monthly mean and deviation from average seasonal value)
- a population count
- a terrain ruggedness index
- the district size

- the share of cropland use
- the share of pasture use

In other words, we collect district-level data on 9 time-varying risk indicators describing 5 different types of risk, and 5 time-invariant risk indicators. The dataset covers 21 out of the 37 countries covered by FEWS NET – Afghanistan, Burkina Faso, Chad, Democratic Republic of the Congo, Ethiopia, Guatemala, Haiti, Kenya, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Somalia, South Sudan, Sudan, Uganda, Republic of Yemen, Zambia, and Zimbabwe. To build our predictive models, we focus on the subset of 15 countries experiencing more than 20 food crises during the observation period – Afghanistan, Chad, Ethiopia, Guatemala, Haiti, Kenya, Malawi, Mali, Niger, Nigeria, Somalia, South Sudan, Sudan, Republic of Yemen, and Zimbabwe. This dataset contains 33,847 monthly observations across 915 districts between July 2009 and February 2020 (Fig. 4.5).

4.5.2 Regression model

Let $v_{k,d,t}$ be the value of time-varying risk indicator k in district d during the month ending on date t and let $v_{l,d}$ be the value of time-invariant risk indicator l in district d. In addition, let $x_{w,d,t}$ be the news factor measuring the proportion of news articles mentioning text feature w and district d during the month ending on date t. To account for news mentions of causes of food insecurity co-occurring with the name of a province or a country, we also introduce $x_{w,p_d,t}$ and $x_{w,c_d,t}$ where p_d and c_d respectively correspond to the province and the country that district d belongs to. Finally, let $y_{d,t}$ be the IPC phase in district d during the quarter ending on date t, such that missing data are filled forward using the latest available data. To predict the IPC phase, we estimate the following panel autoregressive distributed lag (ADL) model:

$$y_{d,t} = a_d + \sum_{m=1}^{6} a_{d,m} y_{d,t-3m} + \sum_{k=1}^{9} \sum_{n=1}^{6} b_{k,d,n} v_{k,d,t-2-n} + \sum_{l=1}^{5} b_{l,d} v_{l,d} + \sum_{w=1}^{167} \sum_{n=1}^{6} b_{w,d,n} x_{w,d,t-2-n} + b_{w,p_d,n} x_{w,p_d,t-2-n} + b_{w,c_d,n} x_{w,c_d,t-2-n}. \quad (4.2)$$

We set: $b_{w,d,n} = b_{w,p_d,n} = b_{w,c_d,n} = 0$ to estimate the baseline model and $b_{k,d,n} = b_{l,d} = 0$ to estimate the news-based model. To measure each model's predictive performance, we first partition the observation period into 10 disjoint folds of equal length. We then successively train the model using a leave-one-out cross-validation strategy, ensuring that observations from each training set occur before those of the corresponding test set. We then report the average cross-validation root-mean-square-error (RMSE) across the 10 folds, both for the full model as well as for each country separately. A Lasso regularization worsens the predictions, increasing the out-of-sample RMSE by 9.9%, 1.4%, and 4.4% for the baseline, news-based, and combined model respectively (Fig 4.5).

In these estimates, each news factor is computed independently of whether a target keyword also appears in an article in which a text feature is mentioned (Fig. 4.3A and 4.3C). However, the presence of a target keyword could indicate that the text is suggesting that a food crisis is already unraveling. As a robustness check, we recompute our news factors after having excluded any article containing a target keyword. The resulting reductions in RMSE of the news-based and combined model relative to the baseline model are equal to 33.8% and 39.1% respectively, which represents a small deterioration of the results compared to our preferred estimates presented in Fig. 4.12.

Finally, we investigate whether the predictive performance of the model described in equation 4.2 changes by incorporating spatial averages of district-level terms accounting for the tendency of food insecurity to be spatially correlated. Let $\tilde{y}_{.,d,.}$ be the spatial average of $y_{.,d,.}$ computed

using the 4 nearest neighbors of district d. $\tilde{x}_{.,d,.}$, $\tilde{v}_{.,d,.}$, and $\tilde{v}_{.,d}$ are defined in a similar fashion. We then estimate the following model:

$$
\begin{aligned}
y_{d,t} = a_d &+ \sum_{m=1}^{6} a_{d,m} y_{d,t-3m} + \sum_{k=1}^{9}\sum_{n=1}^{6} b_{k,d,n} v_{k,d,t-2-n} + \sum_{l=1}^{5} b_{l,d} v_{l,d} \\
&+ \sum_{w=1}^{167}\sum_{n=1}^{6} b_{w,d,n} x_{w,d,t-2-n} + b_{w,p_d,n} x_{w,p_d,t-2-n} + b_{w,c_d,n} x_{w,c_d,t-2-n} \\
&+ \sum_{m=1}^{6} a_{d,m} \tilde{y}_{d,t-3m} + \sum_{k=1}^{9}\sum_{n=1}^{6} b_{k,d,n} \tilde{v}_{k,d,t-2-n} + \sum_{l=1}^{5} b_{l,d} \tilde{v}_{l,d} \\
&+ \sum_{w=1}^{167}\sum_{n=1}^{6} b_{w,d,n} \tilde{x}_{w,d,t-2-n}.
\end{aligned}
\tag{4.3}
$$

We obtain an out-of-sample RMSE equal to 15%, 9.5%, 8.8% for the baseline, news-based and combined model respectively (Fig. 4.5), which corresponds to a reduction in RMSE of 1.6%, 5% and 3.5% respectively. Since the predictive gains are modest, we choose to keep the model more parsimonious by presenting estimates from equation 4.2 in Fig. 4.12.

4.5.3 Classification of food crisis outbreaks

We define a food crisis outbreak as a sequence of two consecutive periods during which the IPC phase raises to a value of 3 or more while the previous period's phase is smaller or equal to 2. We aim to predict an outcome variable which is equal to 1 when a food crisis outbreak occurs and zero otherwise. We convert each previously estimated model of the IPC phase into a classifier of food crisis outbreaks by introducing a lower threshold l and an upper threshold u. An outbreak is predicted to start in district d during the quarter ending on date t if and only if:

$$
\begin{aligned}
y_{d,t+1} &\geq u \\
y_{d,t} &\geq u \\
y_{d,t-1} &\leq l.
\end{aligned}
\tag{4.4}
$$

Each model's thresholds l and u determine its precision and its recall. By varying l and u from 1 to 5 in increments of size 0.1, we can estimate a model's Pareto front. We then fix all the models' precision to be equal to 80% and we compare their recall values measured on the Pareto front. We obtain threshold values for (l, u) equal to $(2.2, 3.1)$, $(1.9, 2.7)$, and $(2.1, 3.3)$ for the baseline, news-based, and combined model respectively. While we are agnostic about how to balance type I and type II errors, our results show an improvement along the Precision-Recall curves across countries (Fig 4.8).

At the time of publishing the IPC phase, FEWS NET additionally provides a projection of next period's values ("expert model"). In line with our previous analysis, we binarize these expert forecasts to produce a predictive model of food crisis outbreaks in which an outbreak occurs when the IPC phase raises to 3 or more for at least least two consecutive periods. The expert model's precision and recall are equal to 70% and 66% respectively, which represents a degradation compared to the combined model for which we obtained a precision and a recall respectively equal to 80% and 86%.

Competing Interests and Conflict of Interests Disclosures

Dr. Fraiberger reports receipt of grant funding from the World Bank World Development Report 2021: Data for better lives. Dr. Subramanian is a co-founder of Entrupy Inc, Velai Inc, and Gaius Networks Inc and has served as a consultant for the World Bank and the Governance Lab. Mr.

Balashankar is a Ph.D student at New York University, and is also funded in part, by the Google Student Research Advising Program. Dr. Subramanian reports that Velai Inc broadly works in the area of socio-economic predictive models, and Mr. Balashankar holds a small percentage of equity in Velai Inc, for licensed technology through New York University. No other disclosures were reported.

Materials and correspondence

Extended Data

4.6 Results

Next, we demonstrate that news factors help predict variations in food insecurity in fragile states (Fig. 4.12). Following previous research on food insecurity early warning systems [243, 14], we first estimate a panel autoregressive distributed lag (ADL) model to predict the IPC phase using past values of the traditional factors described in Fig. 4.11 along with time-invariant risk factors – population count, district size, terrain ruggedness and agricultural land use share ("baseline model"). We then compare the baseline model's predictive performance to that of the same ADL model in which we substitute traditional factors with news factors ("news-based model"), finding that the news-based model leads to a reduction in out-of-sample root-mean-square error (RMSE) [132] of 34.1% relative to the baseline model (Fig. 4.12A). These results suggest that news factors could serve as a proxy for food insecurity risk factors in regions in which other sources of data are unavailable or outdated. Furthermore, when we incorporate both traditional factors and news factors into the same ADL model ("combined model"), the reduction in RMSE relative to the baseline model reaches 40%, proving that news-based indicators of food insecurity also serve as a complement to traditional risk indicators (supplementary text 4.5.2). While these results show

that news factors improve the prediction of variations in food security at the district level, we find a substantial degree of heterogeneity in predictive gains across countries, ranging from 20.5% for Malawi to 48.4% for Mali, which is in part explained by differences in news coverage (Fig. 4.7). To put these results into perspective, we also demonstrate that news factors specifically help predict the outbreak of a food crisis, which corresponds to the IPC phase raising to a value of 3 or more for at least two consecutive periods, an event of utmost importance to disaster relief organizations deciding when and where to allocate emergency food assistance (supplementary text 4.5.3). By converting each previously estimated model into a binary classifier of a crisis outbreak and fixing its precision at 80%, we find that the combined model's out-of-sample recall reaches 86%, compared to 66% for the news-based model and 54% for the baseline model (Fig. 4.12B). In other words, while the baseline model is able to predict 962 out of the 1,797 food crises observed in our validation set, incorporating news factors helps anticipate 581 additional crises which would have otherwise been missed (Fig. 4.12C). In addition, the combined model predicts 47 out of the 48 crisis outbreaks in which the IPC phase escalated to a level 4 or 5, while the baseline and news-based model only predict 26 and 33 of these outbreaks respectively, indicating that news factors are especially valuable in anticipating the most severe outbreaks. Taken together, these results show that news mentions of causes of food insecurity precede variations in district-level IPC phases and could help dispatch emergency relief up to three months ahead of a food crisis.

4.7 Discussion

While machine learning is often criticized for its lack of transparency [250], our model's predictions can easily be interpreted. Focusing on Somalia, South Sudan and Ethiopia, three of the countries which experienced the highest level of food insecurity in recent decades, we zoom in on specific crisis episodes in our validation set to elicit which news factors help predict the deterioration of the situation. The first episode that we analyze happened in 2011 in Somalia, where

the combination of a drought, rising food prices, forced displacement and a sustained conflict led to the worst famine of the 21st century [278]. In particular, the district of Jamaame evolved from an IPC phase 2 during the first half of 2011 to a phase 4 by July, following intensifying violence in the Southern part of the country. While the proportion of news articles mentioning both Jamaame and terms included in the "conflict and violence" cluster started raising 5 months prior to the change in the IPC phase, the conflict fatality count did not record any death in the district until the summer of 2012, highlighting that news factors capture relevant dimensions of civil insecurity which are missing from traditional conflict indicators (Fig. 4.12D and 4.12G). The second episode that we focus on occurred in 2016 when a fall armyworm spread across 20 countries in Africa, decimating large quantities of crops [145]. The worm was first reported in early 2016, and by September, the proportion of news mentioning both the Yambio county in South Sudan and text features included in the "pests and diseases" cluster had peaked, 4 months prior to the inverted vegetation index peaking, and 5 months ahead of the IPC phase raising from 2 to 3 (Fig. 4.12E and 4.12H). Although pest infestations are indirectly measured through vegetation indices, their damage on crops are typically only reflected in vegetation greenness once the food security of neighboring populations has begun to deteriorate, strengthening the importance of measuring anticipatory signals from the news. Finally, the last episode of our study took place in 2009 in Ethiopia when it experienced one of its driest years of the past 50 years, wreaking havoc on food production [421]. Seasonality-adjusted levels of precipitation in the Majang district were 2.3 standard deviations below their historical average in September 2009 before reverting to their mean at the beginning of 2010. While the prolonged effect of this extreme drought was not well captured by a precipitation index, the proportion of news mentioning both Majang and terms contained in the "weather conditions" cluster started increasing in late 2009 and remained close to its peak until July of 2010 when the IPC phase increased from 1 to 3, suggesting that news indices are also better suited to anticipate a drought-related food crisis (Fig. 4.12F and 4.12I). To quantify the role played by news factors in driving our predictions during these episodes, we re-estimate

the combined model after having removed the cluster of news factors containing terms related to "conflicts and violence", "pests and diseases", and "weather conditions" respectively ("ablated models"). For all 3 episodes, we find that the combined model is able to accurately anticipate the change in the IPC phase whereas both the ablated model and the baseline model fail to predict it (Fig. 4.12G-4.12I and 4.9). In other words, risk factors leading to a food crisis are better anticipated by news indicators than by traditional ones which can be incomplete, delayed or outdated, and our model enables us to explicitly interpret predictions of food crisis outbreaks by linking them to variations in news mentions of the underlying causes of an upcoming outbreak.

Although the drivers of food insecurity are well-known, early warning systems relying on high-frequency measurements of these factors are still lacking. The data-driven approach described in this paper could drastically improve the prediction of food crisis outbreaks up to three months ahead of time using real-time news streams and a predictive model that is simple to interpret and explain to policymakers. Development practitioners working for humanitarian organizations such as the World Food Program could use the predictions of our model to help prioritize the allocation of emergency food assistance across vulnerable regions in a principled way, allowing for a more effective preparedness and a reduction in human suffering when a crisis hits. Early warnings cannot address all of the sources of delay in emergency responses, however it can mitigate it by increasing the cost of inaction for governments and the international community [84]. While our study only focuses on news articles in English, future work incorporating local languages into our framework could potentially improve the predictive performance of our model even further. In addition, development practitioners could extend our model to produce estimates of the IPC phase during periods or in regions in which it is not currently being reported, at a fraction of the cost. Beyond the context of food insecurity, our novel approach for selecting causally grounded news indicators addresses the risk of overfitting when big data and machine learning is being used to predict policy outcomes in data-scarce environments [426, 238, 225], and could be extended to other domains, from disease surveillance to the impact of climate change.

Figures

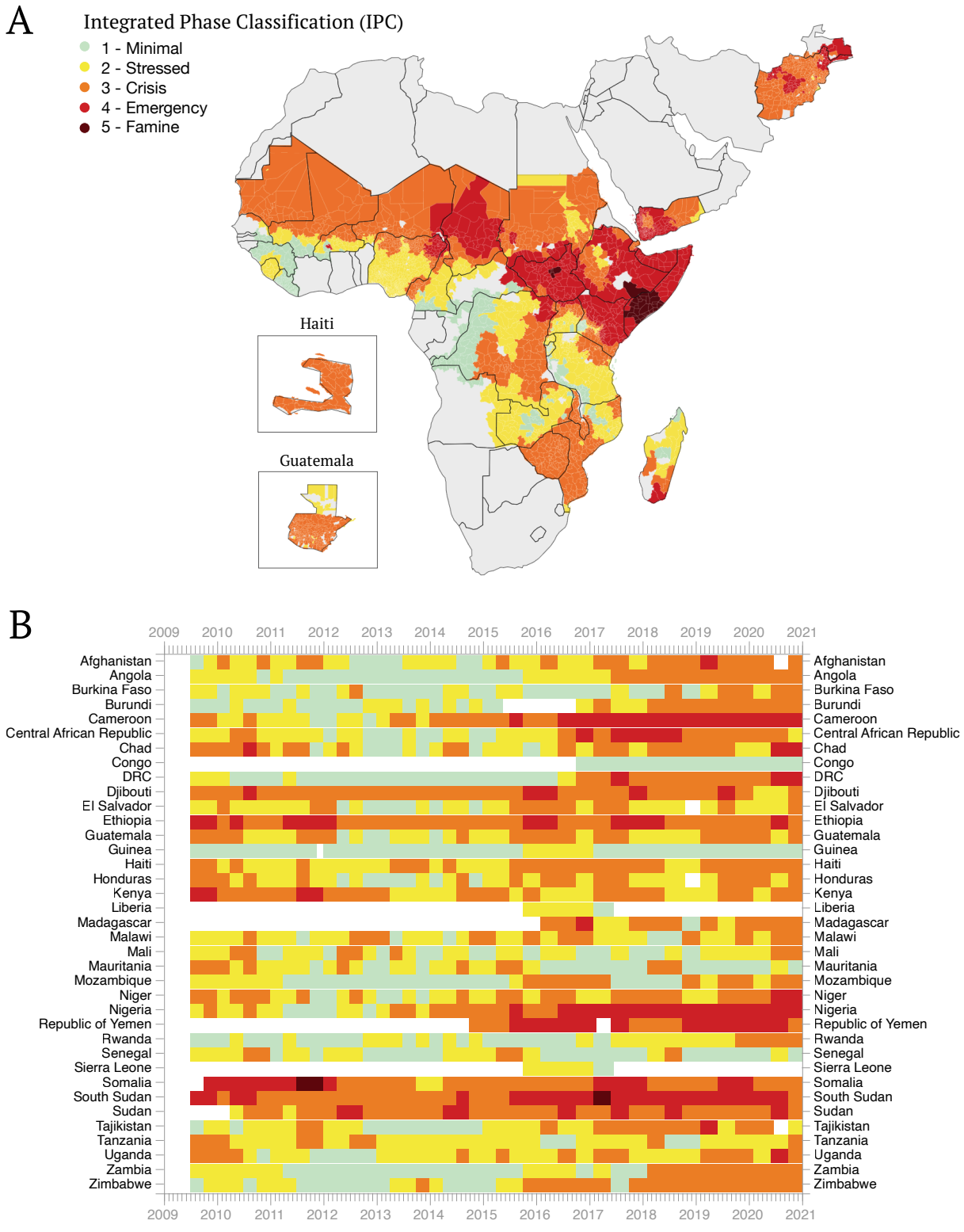


Figure 4.1: Food security dataset. (A) Integrated Phase Classification (IPC) of food security into 5 phases – minimal, stressed, crisis, emergency, and famine – at the district level across the 37 countries covered by the FEWS NET dataset. Each administrative unit is characterized by the maximum IPC phase measured over the period 2009-2020, revealing that food insecurity is geographically clustered. (B) Heatmap showing the maximum value of the IPC phase at the country level during each measurement period.

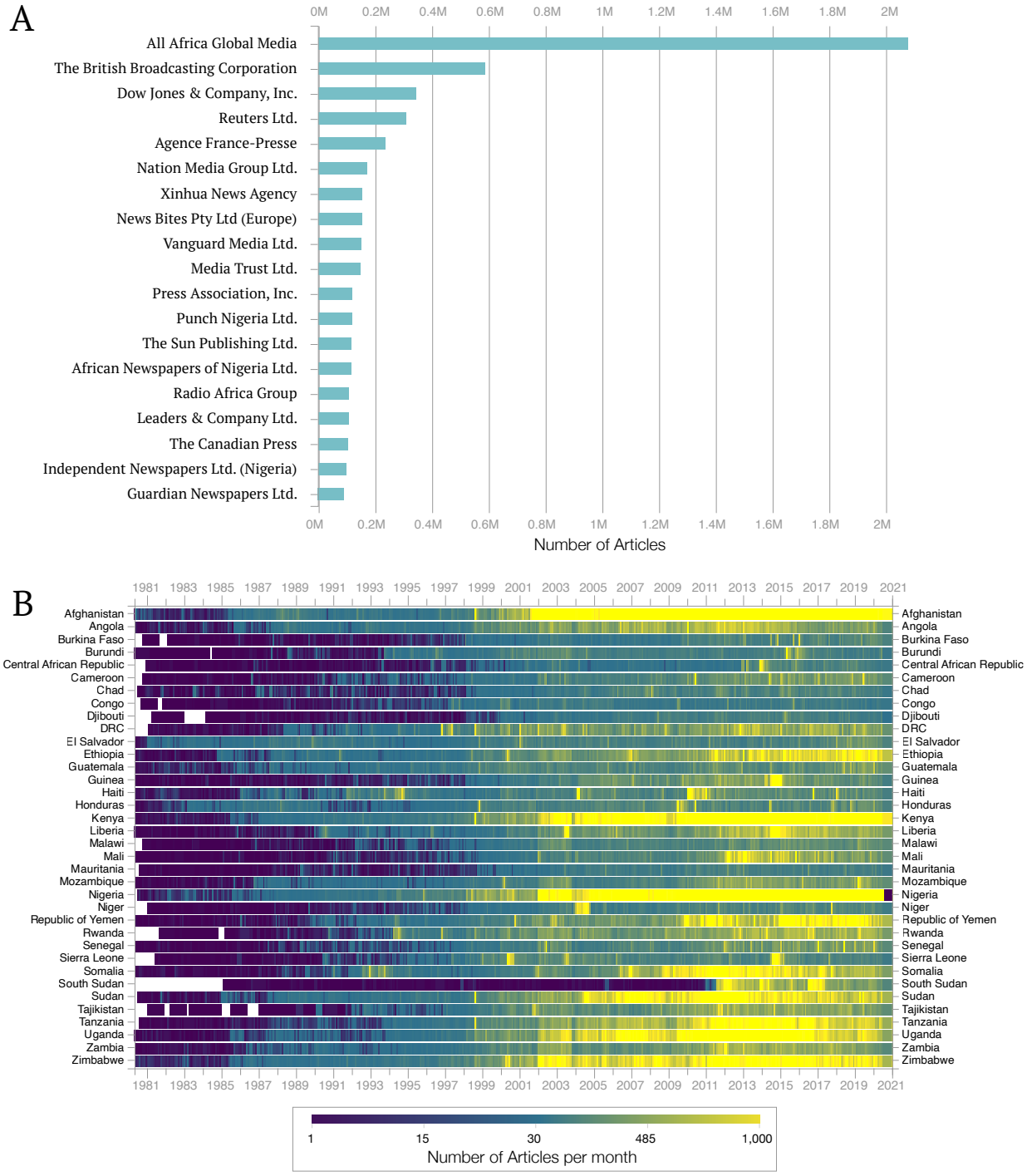


Figure 4.2: Corpus of news articles. (A) The number of news articles grouped by publisher. (B) The number of news articles grouped by month and by country. We use the classification provided by Factiva to establish that an article focuses on a specific country.

A

Target Keyword	Number of Articles
famine	25,637
food insecurity	25,404
malnourished	12,102
malnutrition	10,372
food crisis	10,154
starvation	8,012
hunger crises	6,518
shortage of food	5,482
life-threatening hunger	1,895
lack of food	1,266
scarcity of food	1,058
acute hunger	1,043
dearth of food	891

B

Causal link	Number of Articles
since.c	10,548
see.v	10,142
result.n	9,016
cause.v	8,921
result (in).v	8,731
send.v	8,542
so.c	8,482
resulting.a	8,402
due to.prep	8,231
give rise.v	8,099
lead (to).v	7,601
leave.v	7,011
because.c	6,749
because of.prep	5,932
bring about.v	4,207
ravage.v	4,205
suffer.v	3,621
mean.v	3,047
bring.v	3,037
for.c	2,891
raise.v	2,462
render.v	2,145
bring on.v	2,088
force.v	1,904
reason.n	1,855
wreak.v	1,855
responsible.a	1,521
put.v	1,494
make.v	1,076
precipitate.v	1,047
cause.n	1,043
causative.a	980
sway.v	932
consequence.n	732
consequent.a	630
motivate.v	216
induce.v	108
dictate.v	107
resultant.a	67
consequential.a	29
legacy.n	0

C

Text feature	Number of Articles
rise	1,377,032
conflict	754,194
coup	527,508
terrorism	519,898
corruption	515,890
flee	286,920
refugees	282,894
displaced	253,773
migration	220,003
drought	139,740
tragedy	106,500
climate change	89,612
siege	87,720
looting	71,207
floods	66,570
foreign troops	65,166
pirates	58,269
human rights abuses	57,870
economic crisis	53,800
catastrophe	50,612
repression	47,806
foreign aid	42,924
secession	42,924
land reform	41,067
price rise	39,474
natural disaster	36,672
the offensive	34,698
humanitarian situation	34,453
blockade	34,132
food insecurity	31,608
power struggle	29,977
asylum seekers	25,558
overthrow	23,220
food crisis	23,135
malnourished	23,134
cyclone	22,822
apathy	22,556
farmland	22,355
military junta	21,493
air attack	19,920
military dictatorship	19,916
bombing campaign	19,563
dysfunction	19,332
food assistance	18,324
militia groups	18,119
dehydrated	17,688
epidemics	17,647
dictators	17,620
convoys	17,219
cholera outbreak	17,156
corrupt government	16,853
environmental degradation	16,565
infant mortality	16,418
humanitarian disaster	15,271
totalitarian	14,926
carbon	14,456
civil strife	13,991
pests	13,884
international intervention	12,933
land grab	12,849
gastrointestinal	12,736
d'etat	12,597
rising inflation	11,059
police torture	10,523
jihadist groups	10,231
land seizures	10,151
slave trade	10,112
international terrorists	10,059
warlord	10,045
rebel insurgency	10,030
burning houses	9,967
ecological crisis	9,965
price of food	9,861
terrorist	9,824
politically engineered	9,713
reduced national output	9,562
lack of rains	9,526
internal strife	9,523
years of warfare	9,517
self reliance	9,473
major offensive	9,440
alarming level	9,421
water availability	9,294
transport bottleneck	9,160
inadequate rainfall	9,140
hunger crises	9,114
aid appeal	9,099
rinderpest	9,047
land invasions	9,037
brutal government	8,831
without international aid	8,760
collapse of government	8,753
massive starvation	8,588
locusts	8,566
mass hunger	8,541
greenhouse gases	8,527
mayhem	8,504
rising food prices	7,913
call for donations	7,840
withheld relief	7,748
brain drain	7,744
potato blight	7,733
clans	7,491
gangs of bandits	7,490
water distribution shortages	7,384
lack of authority	7,257
continued strife	7,199
infrastructure damage	7,039
harvest decline	6,871
pushing peasants off	6,712
mismanagement	6,645
scanty rainfall	6,580
forests destroyed	6,576
toll on livestock	6,523
makeshift camps	6,457
cycle of poverty	6,435
wreaked havoc	6,419
failed crops	6,381
harvests are devastated	6,337
bad harvests	6,205
stolen food aid	6,058
destructive pattern	5,862
slashed export	5,801
regimes were toppled	5,783
land degradation	5,770
increased external debt	5,717
man-made disaster	5,688
prolonged fighting	5,625
cattle death	5,608
disrupted trade	5,578
prolonged dry spell	5,545
rival warlords	5,373
authoritarian	5,106
shortage of rains	4,801
economic impoverishment	4,510
severe rains	4,435
failed rains	4,429
life-threatening hunger	4,314
violent suppression	4,155
international embargo	4,134
disruption to farming	3,690
restricted humanitarian access	3,627
clan warfare	3,623
collapsing economy	3,459
climatic hazards	3,446
weather extremes	3,435
poor soil quality	3,310
acute hunger	2,997
lack of agricultural infrastructure	2,877
lack of roads	2,772
reduced imports	2,768
population crisis	2,765
cattle plague	2,643
continued deterioration	2,593
clan battle	2,370
aid workers died	2,154
international alarm	2,102
devastated the economy	1,853
anti-western policies	1,837
lack of alternatives	1,768
livestock had died	1,668
abnormally low rainfall	1,626
oppressive regimes	1,564
lack of cultivation	1,551
restricted relief flights	1,363
unable to sow	1,136
civilians uprooted	1,087

D

Author & Title

" Brinkman, CS Hendrix",Food Insecurity and Violent Conflict: Causes
" Clay, BK Holcomb",Politics and the Ethiopian famine: 1984-1985
" Ilaboya, E Atikpo, FE Omofuma…","Causes, effects and way forward to food insecurity"
" Ilaboya, E Atikpo, FE Omofuma…","Causes, effects and way forward to food insecurity"
" Legwegoh, EDG Fraser",Food crisis or chronic poverty: metanarratives of food insecurity in Sub-Saharan Africa
" Smith, AE El Obeid, HH Jensen",The geography and causes of food insecurity in developing countries
"Borji, M Moradi, M Otaghi…","Relationship between nutritional status, food insecurity, and causes of hospitalization of children with infectious diseases"
"Clapp, MJ Cohen",The global food crisis: Governance challenges and opportunities
"Eneyew, W Bekele",Causes of household food insecurity in Wolayta: Southern Ethiopia
"Headey, S Fan",Reflections on the global food crisis: How did it happen? How has it hurt? And how can we prevent the next one?
"Hesselberg, JA Yaro",An assessment of the extent and causes of food insecurity in northern Ghana using a livelihood vulnerability framework
"Kidane, ZG Alemu, G Kundhlande","Causes of household food insecurity in Koredegaga peasant association, Oromiya zone, Ethiopia"
"Meng, N Qian, P Yared","The institutional causes of China's Great Famine, 1959–1961"
"Messer, MJ Cohen","Conflict, food insecurity and globalization"
"Mokyr, CO Gráda",What do people die of during famines: the Great Irish Famine in comparative perspective
"Musemwa, V Muchenje, A Mushunje, F Aghdasi…","Household food insecurity in the poorest province of South Africa: level, causes and coping strategies"
"Okori, J Obua, V Baryamureeba",Famine disaster causes and management based on local community's perception in Northern Uganda
"Owusu, F Ng'ambi",Structural damage: The causes and consequences of Malawi's food crisis
"Von Braun, T Teklu, P Webb","Famine in Africa: Causes, responses, and prevention"
Abdalla,Causes of food insecurity in Southern Africa: An assessment
Alex De Waal et al., New Variant Famine: Aids and Food Crisis in Southern Africa.
Alex De Waal, War and Famine in Africa
Alinovi et al., Measuring household resilience to food insecurity: application to Palestinian households
Amartya Sen, Poverty and Famines..an Essay on Entitlement..and Deprivation.
Babatunde, Determinants of vulnerability to food insecurity: A genderbased analysis of farming households in Nigeria
Bail,Understanding the causes of African famine
Barrett, Measuring food insecurity
Bastiaan T. Heijmansa Et Al, Persistent Epigenetic Differences Associated With,prenatal Exposure to Famine in Humans.
Battisti et al., Historical warnings of future food insecurity with unprecedented seasonal heat
Borch, Food security and food insecurity in Europe: An analysis of the academic discourse (1975–2013)
Bowbrick,Rejoinder: an untenable hypothesis on the causes of famine
Bowbrick,The causes of famine: a refutation of Professor Sen's theory
Brookings Institute, Somalia: Drought + Conflict = Famine?
Bryceson, Food insecurity and the social division of labour in Tanzania, 1919-85
Campbell, Food insecurity: a nutritional outcome or a predictor variable?
Chand,"The global food crisis: causes, severity and outlook"
Daniel Maxwell et al., The 2011 Somalia Famine: Context, Causes, and Complications
Deng, Famine in the Sudan: causes, preparedness and response: a political, social and economic analysis of the 1998 Bahr el Ghazal
Devereux, Food insecurity in Ethiopia
Donald Curtis et al., Preventing Famine.policies and Prospects.for Africa.
Eneyew et al., Causes of household food insecurity in Wolayta: Southern Ethiopia
Fawole, E Ilbasmis, B Ozkan","Food insecurity in Africa in terms of causes, effects and solutions: A case study of Nigeria
Foley, An ecological analysis of factors associated with food insecurity in South Australia, 2002–7
Gebre-egziabher Kirosa et al., War, Famine and Excess Child Mortality.in Africa: the Role of Parental Education.
Golay,The food crisis and food security: Towards a new world food order?
Gonzalez,"The global food crisis: law, policy, and the elusive quest for justice"
Hart, Exploring definitions of food insecurity and vulnerability: time to refocus assessments
Hendriks, The food security continuum: a novel tool for understanding food insecurity as a range of experiences
Hendrix, Food insecurity and conflict dynamics: Causal linkages and complex feedbacks
Humanitarian Practice Network, Humanitarian Exchange
Jaeger,The causes of Africa's food crisis
Jaeger,The causes of Africa's food crisis
Jane Corbett, Famine and Household Coping Strategies
Jeffrey Clark, Famine in Somalia and The.international Response:.collective Failure
Jo Allen Fair., The Body Politic, the Bodies of Women, and the Politics of Famine in U.s. Television Coverage of Famine in the Horn of Africa,
John Seaman, Famine Mortality in Africa
John Wagner Givens, An Analysis Of..frank Dikötter's.mao's Great Famine.
Joseph C. Miller, The Significance of Drought, Disease And.famine in the Agriculturally Marginal.zones of West-central Africa
K Darkoh,The underlying causes of the food crisis in Africa
K Darkoh,The underlying causes of the food crisis in Africa
Keen,The benefits of famine: a political economy of famine and relief in Southwestern Sudan 1983-89
Lisa Cliggett, Grains from Grass: Aging, Gender and Famine in Rural Africa
Mamadou Baro et al., Persistent Hunger:.perspectives On.vulnerability, Famine,.and Food Security In.sub-saharan Africa
Mark Duffield, War and Famine in Africa
McPherson,The global food crisis: causes and solutions
Mellor, S Gavian","Famine: Causes, prevention, and relief
Menon,Famine in Malawi: Causes and consequences
Michael Watts, Entitlements or Empowerment Famine and Starvation in Africa.
Misselhorn,"Food Insecurity in Southern Africa: Causes and emerging response options from evidence at regional, provincial and local scales
Misselhorn,What drives food insecurity in southern Africa? A meta-analysis of household economy studies
Noah Zerbe, Feeding the Famine? American Food Aid and The.gmo Debate in Southern Africa
Olsson,"On the causes of famine: drought, desertification and market failure in the Sudan"
Otaha, Food insecurity in Nigeria: Way forward
Perez-Escamilla, Food insecurity measurement and indicators
Peter Singer, Famine, Affluence, and Morality.
Ramakrishna, An empirical analysis of food insecurity in Ethiopia: the case of North Wello
Rosset,Food sovereignty and the contemporary food crisis
Rubey,Malawi's food crisis: Causes and solutions
Sen,The causes of famine: a reply
Siamwalla et al, Food insecurity in developing countries
Singh,"Global food crisis: magnitude, causes and policy measures"
Stephan Haggard et al., Hunger And.human Rights:.the Politics of Famine.in North Korea
Stephen Devereux, Famine in the Twentieth Century
Stephen Devereux., Why Does Famine Persist in Africa?
Swift,Understanding and preventing famine and famine mortality
Undp Somalia, Somalia Drought.impact & Needs.assessment.volume I..synthesis Report.
Walker,Famine in Africa—the real causes and possible solutions
Webb, Addressing the" In" in food insecurity
Weiser, Conceptual framework for understanding the bidirectional links between food insecurity and HIV/AIDS
Welderufael, Determinants of households vulnerability to food insecurity in Ethiopia: econometric analysis of rural and urban households
Wiggins,Lessons from the current food crisis in Southern Africa
Wiggins,Southern Africa's food and humanitarian crisis of 2001–04: causes and lessons
Wolfe, Understanding the experience of food insecurity by elders suggests ways to improve its measurement

Figure 4.3: Semantic-frame parsing. (A) The 13 target keywords used to select semantic frames related to food insecurity along with the number of news articles in which the selected frames appear. To account for possible inflections, we use the Porter stemming algorithm on each word token and we select from our news corpus semantic frames matching the root words. (B) The 41 causal links obtained from the FrameNet lexical database used to select relevant semantic frames, along with the number of news articles in which the selected frames appear. (C) The 167 text features used in our predictive model along with the number of news articles in which they appear. (D) The 93 books and peer-reviewed studies on which we also run our semantic parser.

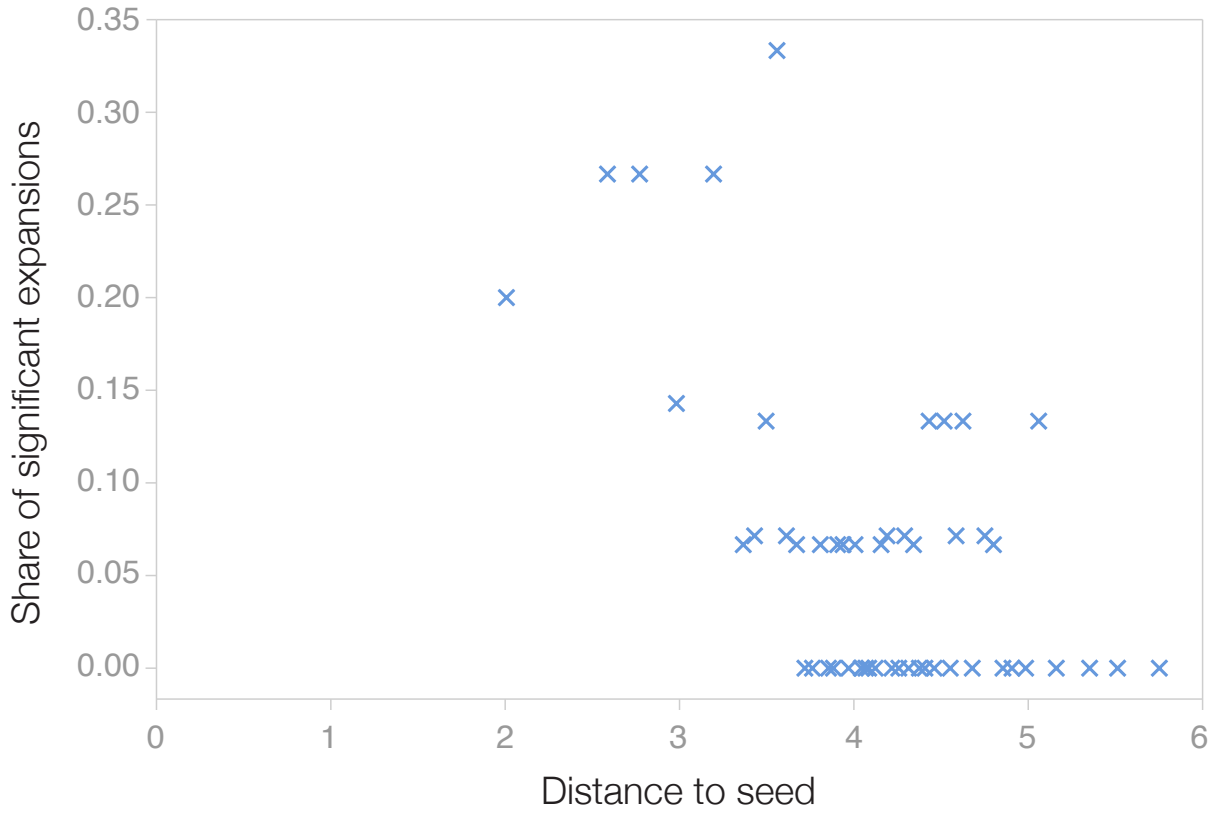


Figure 4.4: Keyword expansion. Starting from the 101 seeds obtained by semantic-frame parsing and passing the Granger causality test, we find the 738 candidate features mentioned in the news and with a word mover's distance to a seed smaller than 6. After ranking candidate features by increasing distance to a seed and partitioning them into 50 groups of equal size, we report the proportion of candidate features within each group passing the Granger causality test (y-axis) and the average distance to a seed within each group (x-axis). As the distance to a seed gets close to 6, the proportion of candidate features predicting the IPC phase approaches zero, providing support to our choice of exploring the space of semantic neighbors up to a distance of 6.

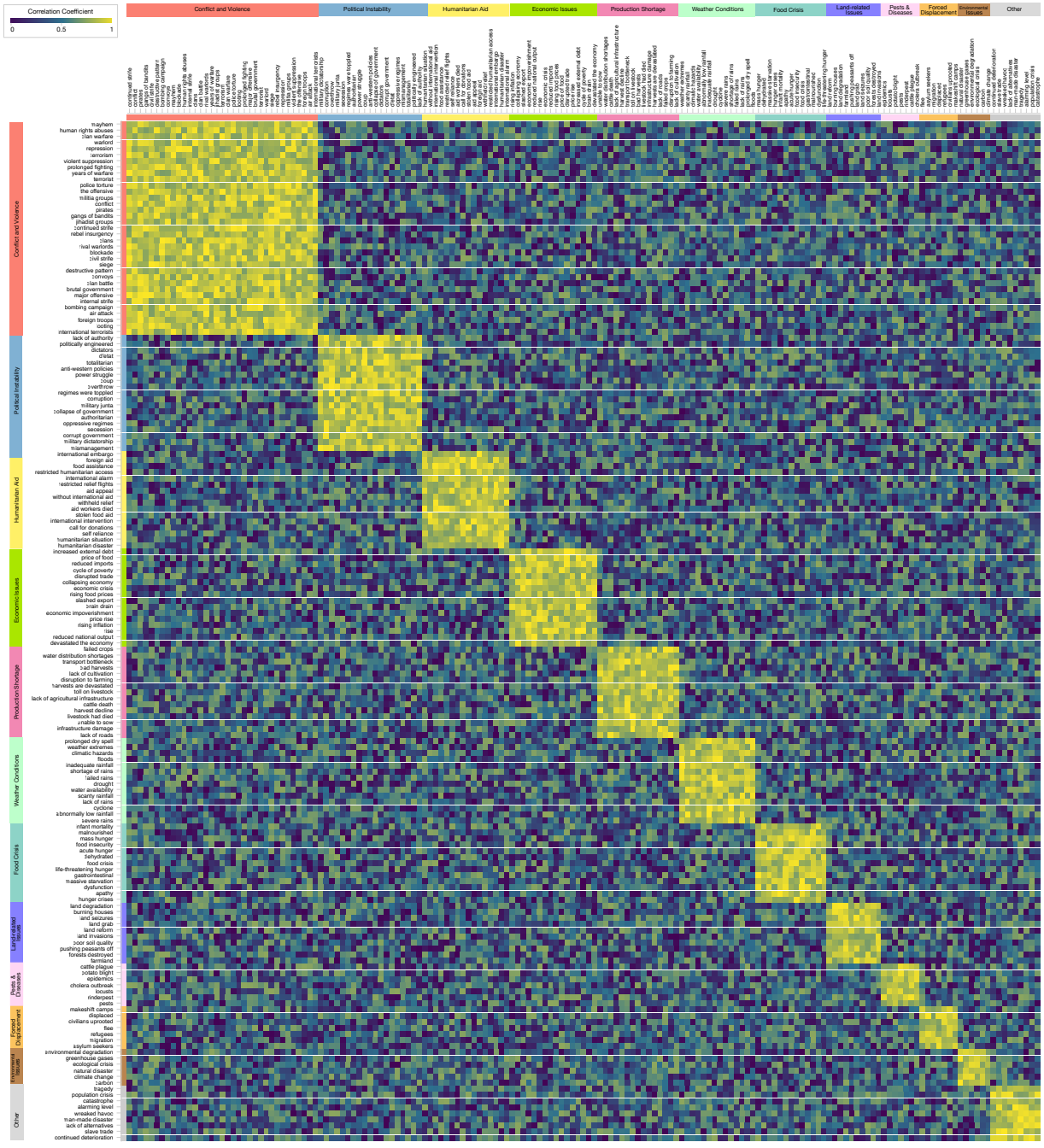


Figure 4.5: Clustering text features. Pairwise correlation between news factors over the period 1980-2020, showing an average correlation between news factors in the same cluster about twice as high as that of factors belonging to different clusters (69.9% versus 34.9%), which provides support to our choice of clustering of our text features into 12 semantically distinct clusters.

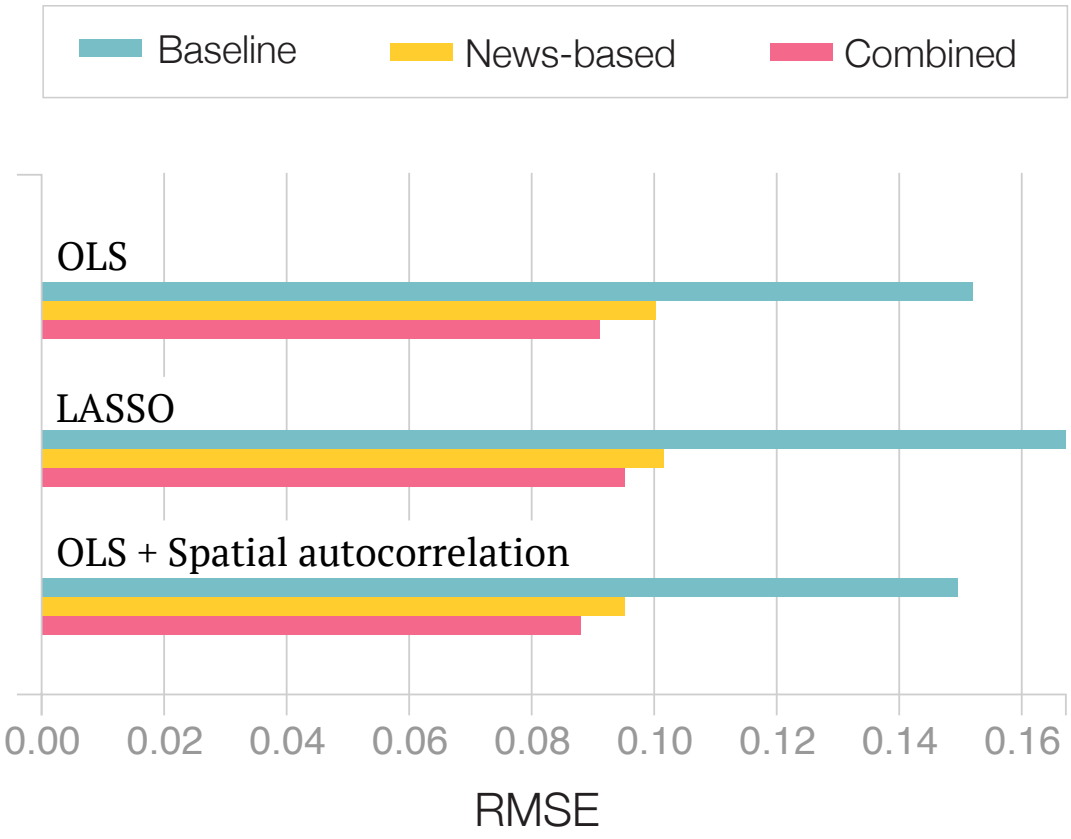


Figure 4.6: Alternative specifications. We first compare the OLS estimates of equation 4.2 shown in Fig.4.12A with estimates of the same model using lasso regularization, showing that it leads to a degradation of the out-of-sample RMSE. We then demonstrate that the model described by equation 4.3 which incorporates spatial averages of district-level terms leads to a small reduction of the out-of-sample RMSE. Since the predictive gains are modest, we choose equation 4.2 as our main specification to keep the model more parsimonious.

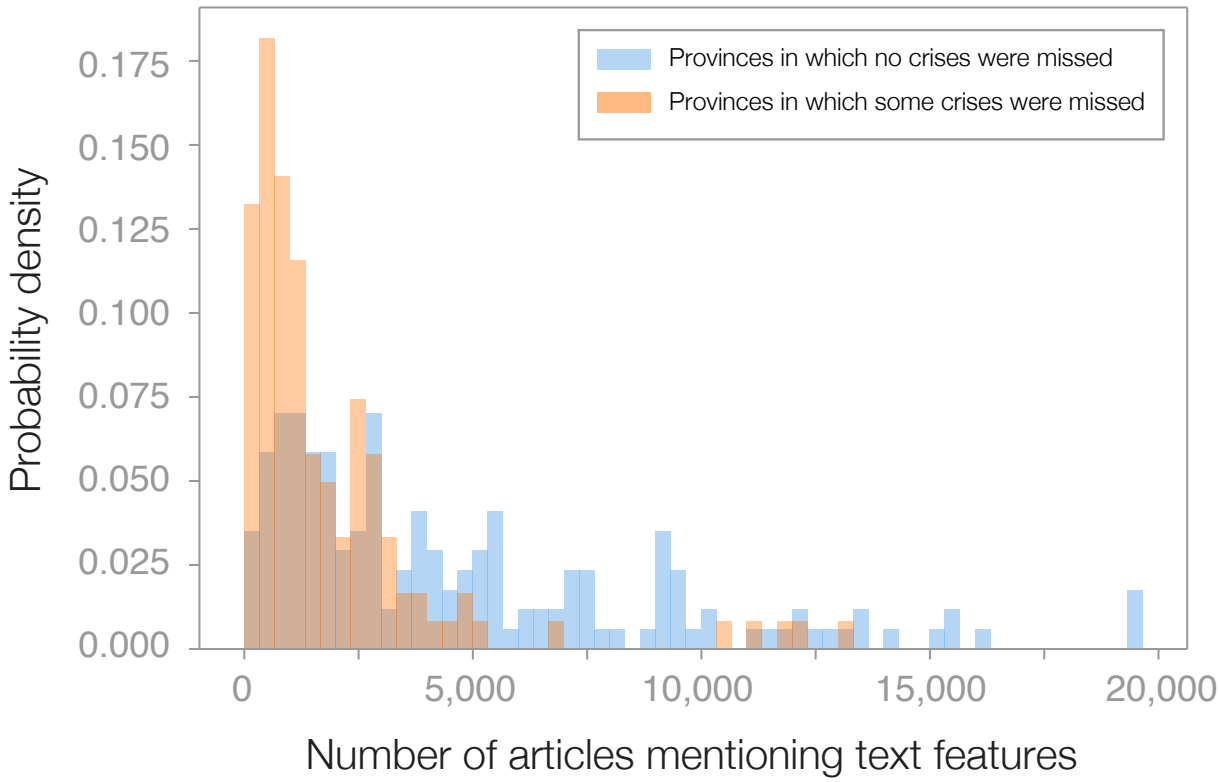


Figure 4.7: News coverage and predictive performance. Distribution of the number of news articles mentioning text features across administrative units of level 1 ("provinces"), separating between provinces in which the combined model predicts all the crisis outbreaks (blue) from those in which it fails to predict at least one crisis (orange), which reveals that provinces in which the combined model fails to predict some crisis outbreaks have lower news coverage that those in which the model predicts all of them.

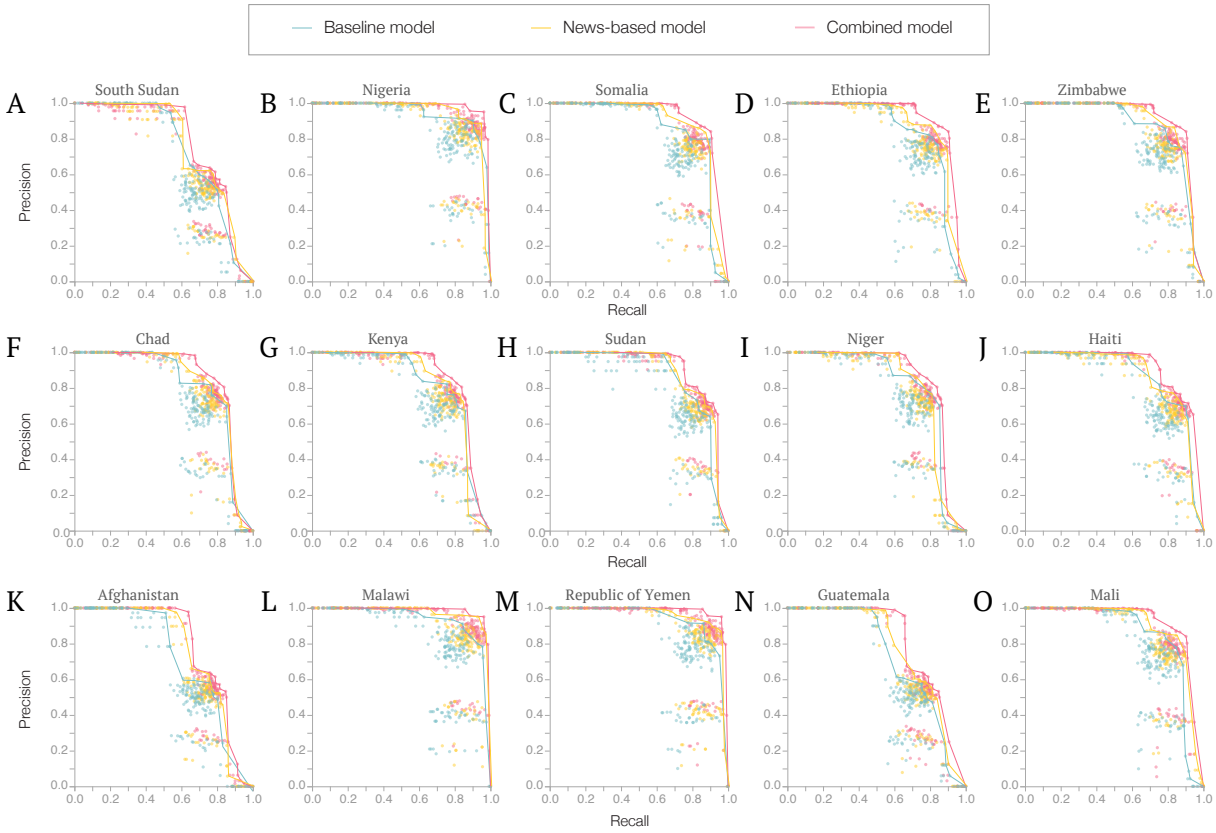


Figure 4.8: Precision-Recall curves. We show the same precision-recall curves as the one described in Fig.4.12B, after having split the evaluation set by country, which indicates that the combined model also outperforms both the news-based and the baseline model at the country level.

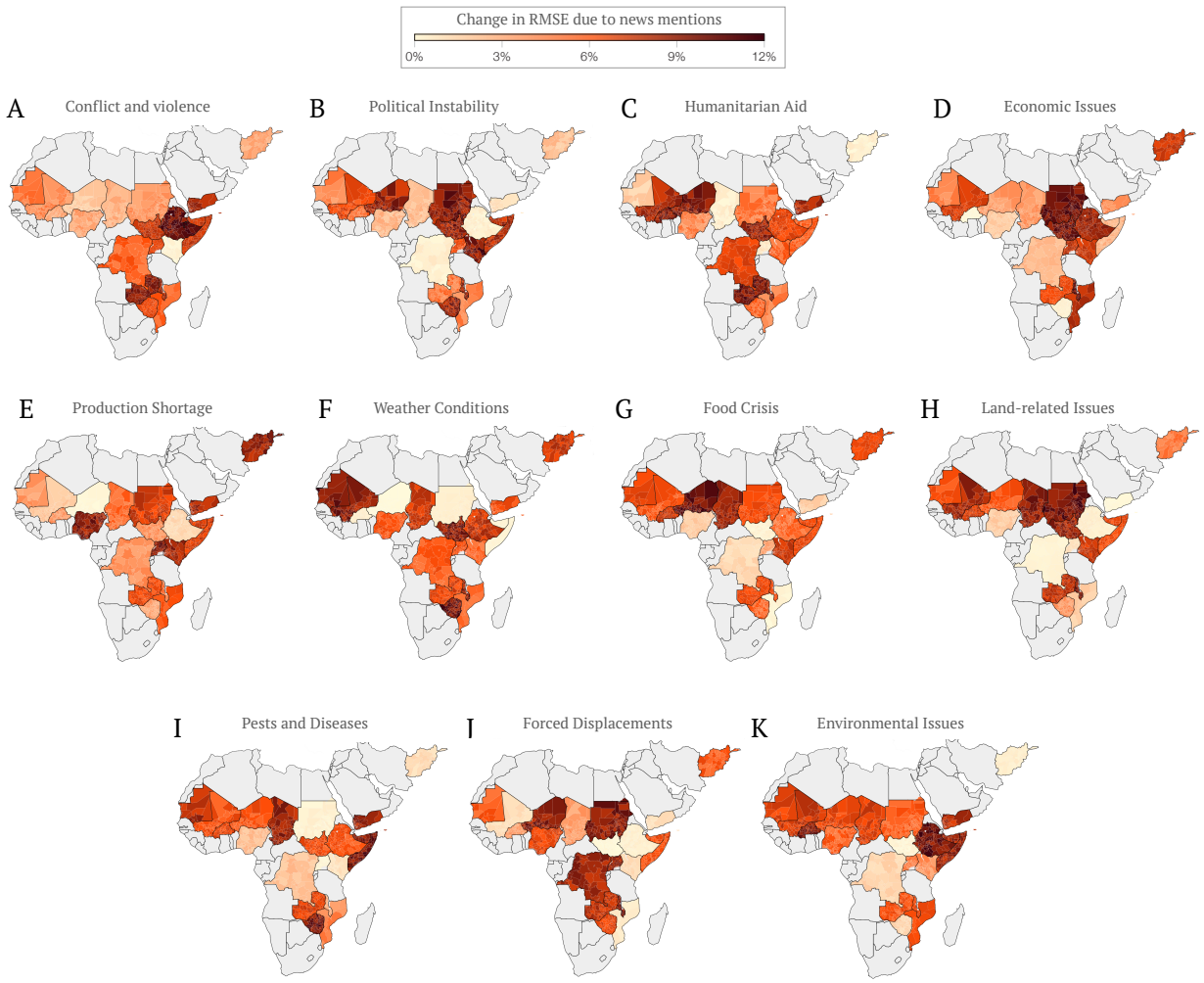


Figure 4.9: Ablated models. We re-estimate the combined model by removing each cluster of news factors ("ablated model"). We report the district-level increase in RMSE of each ablated model relative to the combined model (A-K), allowing us to identify the regions in which each cluster of news factors provides the highest contribution to the prediction.

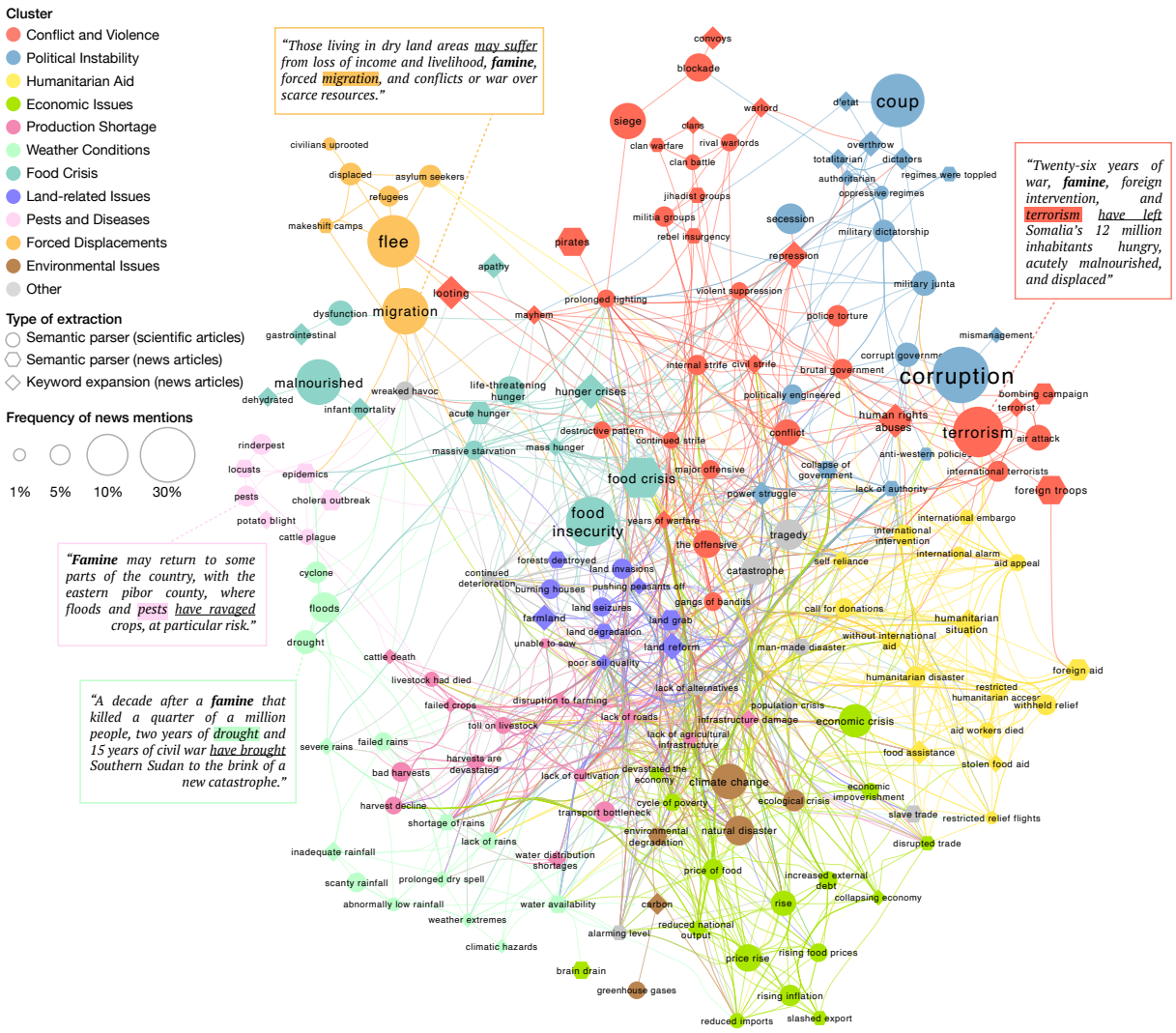


Figure 4.10: Uncovering mentions of causes of food insecurity. Starting from a handcrafted list of 13 target keywords related to food insecurity, we use a frame-semantic parser to extract from scientific (circles) and news (hexagons) articles a seed list of causes of food insecurity ("text features") mentioned in the same semantic frame as a target keyword [24, 394]. Each box contains an example of a sentence in which the parser detects a text feature (highlighted in color) mentioned in the same semantic frame as the target keyword "*famine*" (in bold) and a causal link (underlined). We expand this seed list by collecting text features from news articles (diamond) that are semantically similar to a seed according to their word mover's distance [289, 234]. Text features for which the proportion of monthly local news mentions fails to predict the IPC classification of food security are discarded, leading to a final set of 167 features grouped into 12 clusters based on their semantic similarity and mapped onto a network. A node's size is proportional to its text feature's frequency in news articles mentioning target keywords, and an edge's width encodes the semantic proximity between its end nodes text features. A force-directed algorithm determines each node's position, leading nodes representing semantically similar text features to appear close to one another.

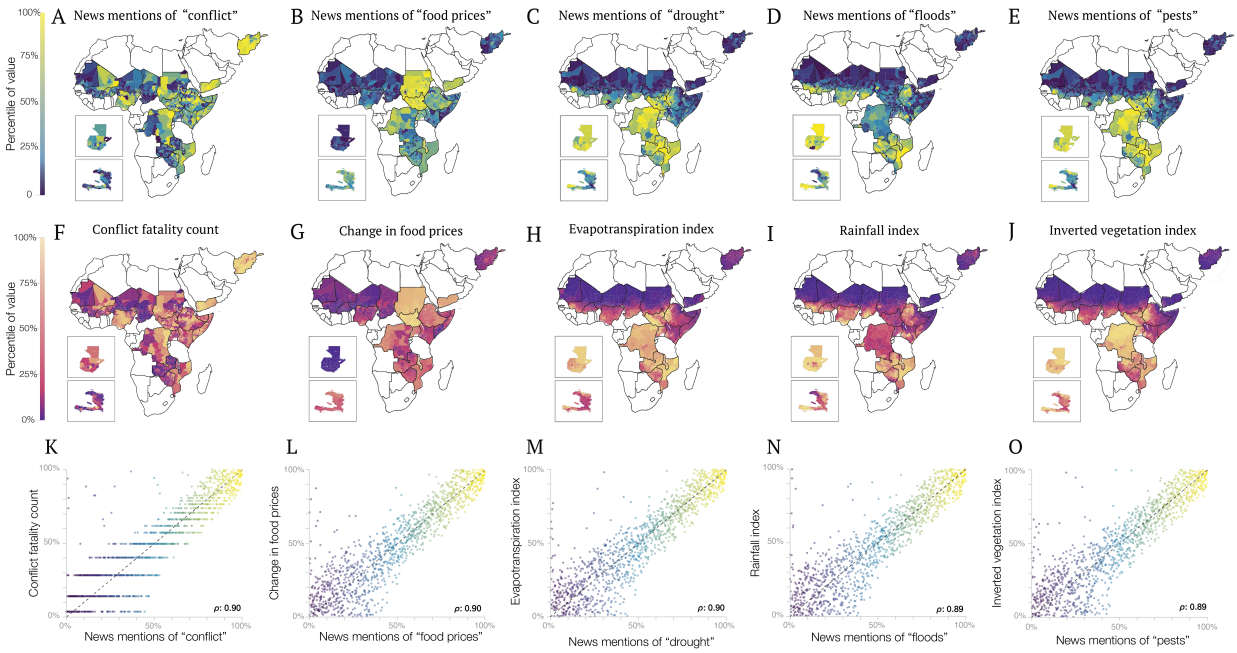


Figure 4.11: Validating news-based indicators of food insecurity. We demonstrate that there exists a strong cross-sectional relationship between news mentions (A-E) and traditional measures (F-J) of causes of food insecurity. We use a comprehensive dataset of traditional measures of food insecurity risk factors ("traditional factors") across 21 fragile states during the period 2009-2020 – a conflict fatality count, the change in food prices, an evapotranspiration index, a rainfall index, and an inverted vegetation index – summarizing each district by the maximum monthly value of each traditional factor during the observation period. To uncover the text feature most closely related to each traditional factor, we first summarize each district by the maximum monthly proportion of local news articles mentioning each text feature ("news factor"). We then associate with each traditional factor the news factor with which it has the highest Spearman correlation across district. (K-O) A scatter plot of a traditional factor (y-axis) and its associated news factor (x-axis) across districts reveals a high Spearman correlation ($r_S > 0.89$). All the values are reported in percentiles.

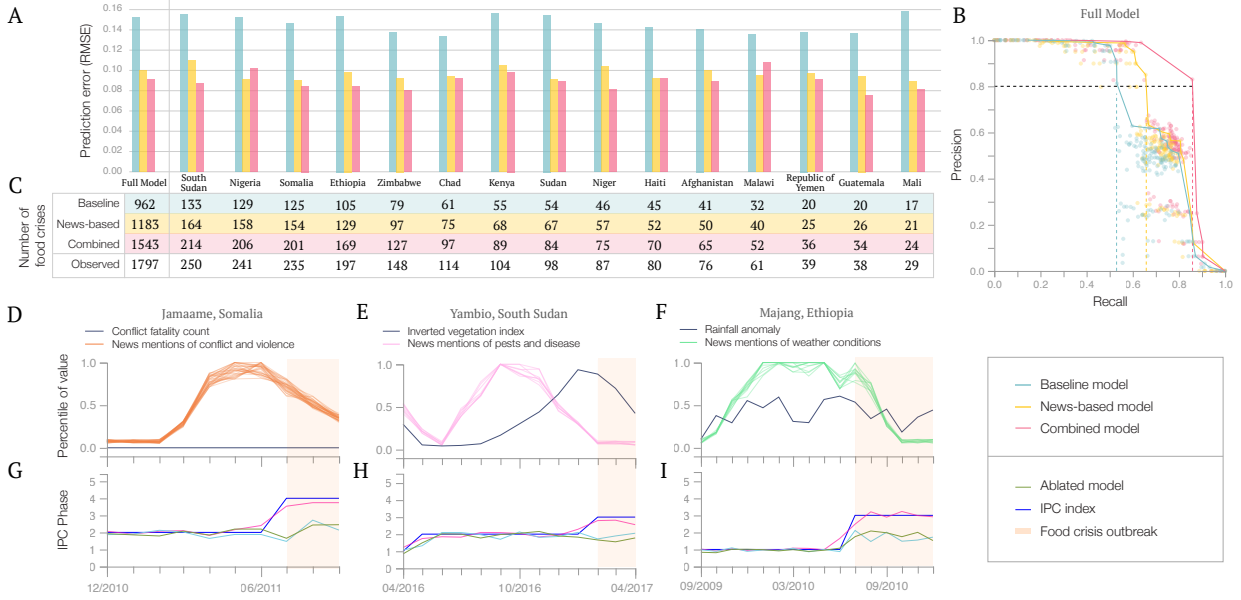


Figure 4.12: Predicting food insecurity. (A) We show that the monthly proportion of local news articles mentioning a text feature ("news factor") helps predict the IPC classification of food security at the district level across 15 fragile states during the period 2009-2020. We estimate a panel autoregressive distributed lag model of the IPC phase on past values of traditional risk factors ("baseline model", turquoise bars), news factors ("news-based model", yellow bars), and both sets of factors ("combined model", pink bars). We report the average root-mean-square error (y-axis) over 10 cross-validated periods, which reveals that including news factors leads to an average reduction in prediction error of 40% relative to the baseline model, with gains ranging from 20.5% for Malawi to 48.4% for Mali. (B) We turn each previously estimated model into a classifier of the outbreak of a food crisis, characterized by the IPC phase raising to a value of 3 or more for at least two consecutive periods. By varying the classification threshold, we construct a series of classifiers (dots) with different precision (y-axis) and recall (x-axis), allowing us to uncover each model's Pareto front (full lines). We then choose each model's threshold such that its precision is equal to 80% (black dotted line), finding that the combined model's recall reaches 86%, compared to 66% for the news-based model and 54% for the baseline model (colored dotted lines). (C) Number of crisis outbreaks observed in the validation set (white row) and predicted by the baseline (turquoise row), news-based (yellow row) and combined (pink row) model at a fixed precision of 80%. (D-F) To elicit the role played by news factors in driving our predictions, we zoom in on 3 crisis episodes in the validation set during which news mentions of causes of food insecurity included in the "conflict and violence" (orange lines), "pests and diseases" (pink lines), and "weather conditions" cluster (green lines) would have helped anticipate the deterioration of the situation. For each episode, we report each text feature's proportion of monthly local news mentions and the most closely related traditional risk indicator (black line). All the values are reported in percentiles. (G-I) We also report the time series of the IPC phase (blue line), and its predicted value using the baseline (turquoise line), ablated (khaki line), and combined (red line) model. While risk factors measured with traditional data fail to provide a warning signal in a timely fashion, news factors peak prior to each crisis outbreak (mauve shaded area), leading the combined model to accurately predict the outbreak whereas the baseline and ablated models fail to predict it.

5 Heterogeneous Granger Causal Factors from News Streams

5.1 Introduction

Contextual embedding models [86] have managed to produce effective representations of words, achieving state-of-the-art performance on a range of NLP tasks. In this paper, we consider a specific task of predicting variations in stock prices based on word relationships extracted from news streams. Existing word embedding techniques are not suited to learn relationships between words appearing in different documents and contexts [239]. Existing work on stock price prediction using news have typically relied on extracting features from financial news [108, 165], or sentiments expressed on Twitter [272, 348, 34], or by focusing on features present in a single document [204, 382]. However, relationships between events affecting stock prices can be quite complex, and their mentions can be spread across multiple documents. For instance, market volatility is known to be triggered by recessions; this relationship may be reflected with a spike in the frequency of the word "recession" followed by a spike in the frequency of the word "volatility" a *few weeks later*. Existing methods are not well-equipped to deal with these cases.

This paper aims to uncover latent relationships between words describing events in news streams, allowing us to unveil *hidden* links between events spread across time, and integrate them into a news-based predictive model for stock prices. We propose the *Predictive Causal Graphs*

(PCG), a framework allowing us to detect latent relationships between words when such relationships are not directly observed. PCG differs from existing relationship extraction [78] and representational frameworks [290] across two dimensions. First, PCG identifies unsupervised causal relationships based on consistent time series prediction instead of association, allowing us to uncover paths of *influence* between news items. Second, PCG finds inter-topic influence relationships outside the "context" or the confines of a single document. Construction of PCG naturally leads to *news-dependent* predictive models for numerous variables, like stock prices.

We construct PCG by identifying Granger causal pairs of words [157] and combining them to form a network of words using the Lasso Granger method [16]. A directed edge in the network therefore represents a potential influence between words. While predictive causality is not true causality [280], identification of predictive causal factors which prove to be relevant predictors over long periods of time provides guidance for future causal inference studies. We achieve this consistency by proposing a framework for *Longitudinal Predictive Causal Factor* identification based on methods of honest estimation [21]. Here, we first estimate a universe of predictive causal factors on a relatively long time series and then identify time-varying predictive causal factors based on constrained estimation on multiple smaller time series. We also augment our model with an orthogonal spike correction ARIMA [49] model, allowing us to overcome the drawback of slow recovery in smaller time series.

We constructed PCG from news streams of around 700, 000 articles from Google News API and New York Times spread across over 6 years and evaluated it to extract features for stock price predictions. We obtained *two orders lower* prediction error compared to a similar semantic causal graph-based method [208]. The longitudinal PCG provided insights into the variation in importance of the predictive causal factors over time, while consistently maintaining a low prediction error rate between 1.5-5% in predicting 10 stock prices. Using full text of more than 1.5 million articles of Times of India news archives for over 10 years, we performed a fine-grained qualitative analysis of PCG and validated that 67% of the semantic causation arguments found in

the news text is connected by a direct edge in PCG while the rest were linked by a path of length 2. In summary, PCG provides a powerful framework for identifying predictive causal factors from news streams to accurately predict and interpret price fluctuations.

5.2 Related Work

Online news articles are a popular source for mining real-world events, including extraction of causal relationships. Radinsky and Horvitz [347] proposed a framework to find causal relationships between events to predict future events from News but caters to a small number of events. Causal relationships extracted from news using Granger causality have also been used for predicting variables, such as stock prices [208, 417, 77]. A similar causal relationship generation model has been proposed by [172] to extract causal relationships from natural language text. A similar approach can be observed in [228, 96], whereas CATENA system [293] used a hybrid approach consisting of a rule-based component and a supervised classifier. PCG differs from these approaches as it explores latent inter-topic causal relationships in an unsupervised manner from the entire vocabulary of words and collocated N-grams.

Apart from using causality, there are many other methods explored to extract information from news and are used in time series based forecasting. Amodeo et al. [12] proposed a hybrid model consisting of time-series analysis, to predict future events using the New York Times corpus. FBLG [62] focused on discovering temporal dependency from time series data and applied it to a Twitter dataset mentioning the Haiti earthquake. Similar work by Luo et al. [262] showed correlations between real-world events and time-series data for incident diagnosis in online services. Other similar works like, Trend Analysis Model (TAM) [216] and Temporal-LDA (TM-LDA) [429] model the temporal aspect of topics in social media streams like Twitter. Structured data extraction from news have also been used for stock price prediction using techniques of information retrieval in [92, 444, 90, 58, 91]. Vaca et al. [411] used a collective matrix factorization method to

track emerging, fading and evolving topics in news streams. PCG is inspired by such time series models and leverages the Granger causality detection framework for the trend prediction task.

Deriving true causality from observational studies has been studied extensively. One of the most widely used algorithm is to control for variables which satisfy the backdoor criterion [328]. This however, requires a knowledge of the causal graph and the unconfoundedness assumption that there are no other unobserved confounding variables. While the unconfoundedness assumption is to some extent valid when we analyze all news streams (under the assumption that all significant events are reported), it is still hard to get away from the causal graph requirement. Propensity score based matching aims to control for most confounding variables by using an external method for estimating and controlling for the likelihood of outcomes [314]. More recently, [428] showed that with multiple causal factors, it is possible to leverage the correlation of those multiple causal factors and deconfound using a latent variable model. This setting is similar to the one we consider, and is guaranteed to be truly causal if there is no confounder which links a single cause and the outcome. This assumption is less strict than the unconfoundedness assumption and makes the case for using predictive causality in such scenarios. Another approach taken by [21] estimates heterogeneous treatment effects by honest estimation where the model selection and factor weight estimation is done on two sub-populations of data by extending regression trees.

Our work is motivated by these works and applies methodologies for time series data extracted from news streams. PCG can offer the following benefits for using news for predictive analytics – (1) Detection of influence path, (2) Unsupervised feature extraction, (3) Hypothesis testing for experiment design.

5.3 Predictive Causal Graph

Predictive Causal Graph (PCG) addresses the discovery of *influence* between words that appear in news text. The identification of influence link between words is based on temporal co-variance, that can help answer questions of the form: "Does the appearance of word x influence the appearance of word y after δ days?". The influence of one word on another is determined based on pairwise causal relationships and is computed using the Granger causality test. Following the identification of Granger causal pairs of words, such pairs are combined together to form a network of words, where the directed edges depict potential influence between words. In the final network, an edge or a path between a word pair represents a flow of influence from the source word to the final word and this *influence* depicts an increase in the appearance of the final words when the source word was observed in news data.

Construction of PCG from the raw unstructured news data, finding pairwise causal links and eventually building the influence network involves numerous challenges. In the rest of the section, we discuss the design methodologies used to overcome these challenges and describe some properties of the PCG.

5.3.1 Selecting *Informative* Words:

Only a small percentage of the words appearing in news can be used for meaningful information extraction and analysis [269, 184]. Specifically, we eliminated too frequent (at least once in more than 50% of the days) or too rare (appearing in less than 100 articles) [**manning_raghavan_sch\IeC {\"u}tze_20**
Many common English nouns, adjectives and verbs, whose contribution to semantics is minimal [114] were also removed from the vocabulary. However, named-entities were retained for their newsworthiness and a set of "trigger" words were retained that depict events (e.g. flood, election) using an existing "event trigger" detection algorithm [4]. The vocabulary set was enhanced by adding bigrams that are significantly collocated in the corpus, such as, 'fuel price' and 'prime

minister' etc.

5.3.2 Time-series Representation of Words:

Consider a corpus D of news articles indexed by time t, such that D_t is the collection of news articles published at time t. Each article $d \in D$ is a collection of words W_d, where i^{th} word $w_{d,i} \in W_d$ is drawn from a vocabulary V of size N. The set of articles published at time t can be expressed in terms of the words appearing in the articles as $\{\alpha_1^t, \alpha_2^t, ..., \alpha_N^t\}$, where α_i^t is the sum of frequency of the word $w_i \in V$ across all articles published at time t. α_i^t corresponding to $w_i \in V$ is defined as, $\alpha_i^t = \frac{\mu_i^t}{\sum_{t=1}^{T} \mu_i^t}$ where $\mu_i^t = \sum_{d=1}^{|D_t|} TF(w_{d,i})$. α_i^t is normalized by using the frequency distribution of w_i in the entire time period. $\mathcal{T}(w_i)$ represents the time series of the word w_i, where i varies from 1 to N, the vocabulary size.

5.3.3 Measuring Influence between Words

Given two time-series X and Y, the Granger causality test checks whether the X is more effective in predicting Y, than using just Y and if this holds then the test concludes X "Granger-causes" Y [157]. However, if both X and Y are driven by a common third process with different lags, one might still fail to reject the alternative hypothesis of Granger causality. Hence, in PCG, we explore the possibility of causal links between all word pairs and detect triangulated relations to eliminate the risk of ignoring confounding variables, otherwise not considered in the Granger causality test.

However, constructing PCG using an exhaustive set of word pairs does not scale, as even after using a reduced set of words and including the collocated phrases, the vocabulary size is around $39,000$. One solution to this problem is considering the Lasso Granger method [16] that applies regression to the neighborhood selection problem for any word, given the fact that the best regressor for that variable with the least squared error will have non-zero coefficients only

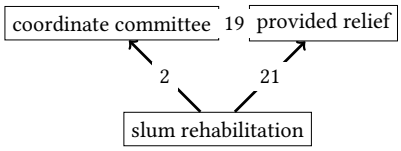


Figure 5.1: PCG highlighting the underlying cause

for the lagged variables in the neighborhood. The Lasso algorithm for linear regression is an incremental algorithm that embodies a method of variable selection [404].

If we define V to be the input vocabulary from the news dataset, N is the vocabulary size, x is the list of all lagged variables (each word is multivariate with a maximum lag of 30 days per word) of the vocabulary, w is the weight vector denoting the influence of each variable, y is the predicted time series variable and λ is a sparsity constraint hyperparameter to be fine-tuned, then minimizing the regression loss below leads to weights that characterize the influential links between words in x that predicts y,

$$\mathbf{w} = argmin \frac{1}{N} \Sigma_{(\mathbf{x},y)\in V} |\mathbf{w}.\mathbf{x} - y|^2 + \lambda ||\mathbf{w}|| \tag{5.1}$$

To set λ, we use the method based on consistent estimation used in [284]. We select the variables that have non-zero co-efficients and choose the best lag for a given variable based on the maximum absolute value of a word's co-efficient. We then, draw an edge from all these words to the predicted word with the annotations of the optimal time lag (in days) and incrementally construct the graph as illustrated in Figure 5.1.

5.3.4 Topic Influence Compression

To arrive at a sparse graphical representation of PCG, we compress the graph based on topics (50 topics in our case). Topics are learned from the original news corpus using unsupervised Latent Dirichlet Allocation (LDA)[41]. Influence is generalized to topic level by calculating the weight of inter-topic influence relationships as a total number of edges between vertices of two topics.

If we define θ_u and θ_v to be two topics in our topic model and $|\theta_u|$ represents the size of topic θ_u, i.e. the number of words in the topic whose topic-probability is greater than a threshold (0.001), then the strength of influence between topics θ_u and θ_v is defined as,

$$\Phi(\theta_u, \theta_v) = \frac{\text{\# Edges between words in } \theta_u \text{ and } \theta_v}{(|\theta_u| \times |\theta_v|)} \tag{5.2}$$

$\Phi(\theta_u, \theta_v)$ is termed as *strong* if its value is in the 99^{th} percentile of Φ for all topics. Any edge in the original PCG is removed if there are no strong topic edges between the corresponding word nodes. This filtered topic graph has only edges between topics which have high influence strength. This combination of inter-document temporal and intra-document distributional similarity is critical to obtaining temporally and semantically consistent predictive causal factors.

5.4 Prediction Models using PCG

In this section, we present three approaches for building prediction models using PCG namely (1) direct estimation using PCG (2) longitudinal prediction which incorporates short term temporal variations and (3) spike augumented prediction which estimates spikes over a longer time window.

5.4.1 Direct Prediction from PCG

One straightforward way of using PCG for prediction modeling is to use the Lasso regression equation used for identifying the predictive causal factors directly. We first adopt this approach by restricting the construction of PCG to the nodes of concern, which significantly speeds up the computation. This inherently ignores any predictive causal factor which only has an indirect link to the outcome node, as theorized by the Granger Causality framework. In this case, we split the data into a contiguous training data, and evaluate on the remaining testing data. If y represents the target stock time series variable and x represents a multivariate vector of all lagged feature

variables, w represents the coefficient weight vector indexed by the feature variable $z \in x$ and time lag m, p, q in days, a represent a bias constant and ϵ_t represents an i.i.d noise variable, then we predict future values of y as follows.

$$y_t = a + \sum_{i=1}^{m} w_{y,i} y_{t-i} + \sum_{z \in x} \sum_{j=p}^{q} w_{z,j} z_{t-j} + \epsilon_t \tag{5.3}$$

5.4.2 Longitudinal Prediction via Honest Estimation

In scenarios where heterogenous causal effects are to be estimated, it is important to adjust by partitioning the time series into subpopulations which vary in the magnitude of the causal effects [21]. In a news stream, this amounts to constructing the word influence networks given a context specified by a time window. This naive extension however can be quite computationally expensive and can limit the speed of inference considerably. However, if the set of potential causal factors are identified over a larger time series, learning their time varying weights over a shorter training period can significantly decrease the computation required.

Hence, we do a two staged honest estimation approach similar to [21]. In the first stage, multiple sets (instead of trees as in [21]) of predictive causal factors $F(Tr_m)$, that provide overall reduction in root mean squared error ($RMSE$), over training data Tr_m, are gathered for model selection through repeated random initialization of regression weights. For any $f \in F(Tr_m)$, we define f to be a set of predictive factors which when trained over data Tr_m to predict future time series values of the target y, achieves $RMSE(Tr_m, f) < \delta$, for a hyperparameter $\delta > 0$.

$$F(Tr_m) = \bigcup\nolimits_f \{RMSE(Tr_m, f) < \delta\} \tag{5.4}$$

From these sets of predictive causal factors $F(Tr_m)$, we choose the set of features $f_{LPC}(Tr_m, Tr_e)$, which gives the least expected root mean squared error on time windows w of length W uniformly sampled from the unseen training data used for estimation Tr_e through cross validation.

This model selection procedure depends on the size of the smaller time windows W used to fine-tune the factors. The size of this time window is a hyperparameter to trade off long-term stability and short-term responsiveness of the predictive causal factors. We chose the time window based of 30 days for our stock price prediction due to prior knowledge that many financial indicators have a monthly cycle and hence responsiveness within that cycle is desired.

$$E(Tr_e, f) = \sqrt{\mathop{\mathbb{E}}_{w \in Tr_e, |w|=W} MSE(w, f)} \tag{5.5}$$

$$f_{LPC}(Tr_m, Tr_e) = min_{f \in F(Tr_m)}\Big(E(Tr_e, f)\Big) \tag{5.6}$$

We then evaluate the model on an unseen time series Te, where the learnt predictive causal factors and their weights are used for inference.

5.4.3 Spike Prediction

One drawback of using a specific time window for estimating the weights of the predictive causal factors is the lack of representative data in the window used for training. This could mean that predicting abrupt drops or spikes in the data would be hard. To overcome this limitation, we train a composite model to predict the residual error from honest estimation by training on differences in consecutive values of the time series. Let $(\Delta y = y_t - y_{t-1}, \Delta f = f_t - f_{t-1})$ denote time series of the differences of the consecutive values of the labels and the feature variables and let $[\Delta y, \Delta f]$ denote the concatenated input variables of the model. We use a multivariate ARIMA model M of order (p, d, q) [49] where p controls the number of time lags included in the model, q denotes the number of times differences are taken between consecutive values and r denotes the time window of the moving average taken to incorporate sudden spikes in values. Let the actual values of the time series of label y be y^*, the predictions of the honest estimation model be $\hat{y}$, a training sample of significantly longer time window Tr_s with $|Tr_s| >> W$, then the composite model is trained to

predict the residuals, $res = y^* - \hat{y} = E(Tr_s, f)$.

$$M = ARIMA(p, d, q) \tag{5.7}$$

$$M.fit(E(Tr_s, f), [\Delta y, \Delta f]) \tag{5.8}$$

$$\hat{res} = M.predict(Tr_e) \tag{5.9}$$

Augmenting this predicted residual ($\hat{res}$) back to $\hat{y}_{Tr_e}$ gives us the spike-corrected estimate $\hat{y_s}$.

$$\hat{y_s} = \hat{y}_{Tr_e} + \hat{res} \tag{5.10}$$

5.5 Results

In this section, we present the results from direction prediction models from PCG, followed by improvement in stock price prediction due to longitudinal and spike prediction from news streams and compare it to a manually tuned semantic causal graph method. We analyze the time varying factors to explain the gains achieved via honest estimation.

5.5.1 Data and Metrics

The news dataset[1] we used for stock price prediction contains news crawled from 2010 to 2013 using Google News APIs and New York Times data from 1989 to 2007. We construct PCG from the time series representation of its 12,804 unigrams and 25,909 bigrams over the entire news corpora of more than 23 years, as well as the 10 stock prices[2] from 2010 to 2012 for training and 2013 as test data for prediction. The prediction is done with varying step sizes (1,3,5), which indicates the time lag between the news data and the day of the predicted stock price in days. The results

[1] https://github.com/dykang/cgraph
[2] https://finance.yahoo.com

shown in Table 5.1 is the root mean squared error (RMSE) in predicted stock value calculated on a 30 day window averaged by moving it by 10 days over the period and directly comparable to our baseline [208]. To evaluate the time-varying factors over a larger time window, we present average monthly cross validation RMSE % sampled over a 4 year window of 2010-13 in Table 5.3. Please note that the results in Table 5.3 are not comparable with [208] as we report a cross validation error over a longer time window.

5.5.2 Prediction Performance of PCG

To evaluate the causal links generated by PCG, we use it to extract features for predicting stock prices using the exact data and prediction setting used in [208] as our baseline.

Baseline: [208] extract relevant variables based on a semantic parser - SEMAFOR [78] by filtering causation related frames from news corpora, topics and sentiments from tweets. To overcome the problem of low recall, they adopt a topic-based knowledge base expansion. This expanded dataset is then used to train a neural reasoning model which generates sequence of cause-effect statements using an attention model where the words are represented using word2vec vectors. [208]'s CGRAPH based forecasting model - C_{best} model uses the top 10 such generated cause features, given the stock name as the effect and apply a vector auto-regressive model on the combined time series of text and historical stock values.

Comparison with the baseline: Compared to the baseline, note that our features and topics were chosen purely based on distributional semantics of the word time series. Once the features are extracted from PCG, we use the past values of stock prices and time series corresponding to the incoming word edges of PCG to predict the future values of the stock prices using the multivariate regression equation used to determine Granger Causality. As compared to their best error, PCG from unigrams, bigrams or both obtain two orders lower error and significantly outperforms C_{best}. The mean absolute error (MAE) for the same set of evaluations is within 0.003 of the RMSE, which indicates that the variance of the errors is also low. We attribute this gain to the flexibility

Table 5.1: 30 day windowed average of stock price prediction error using PCG

Step size	C_{best}	PCG_{uni}	PCG_{bi}	PCG_{both}
1	1.96	0.022	0.023	0.020
3	3.78	0.022	0.023	0.022
5	5.25	0.022	0.023	0.021

Table 5.2: Stock price predictive factors for 2013 in PCG

Stock symbol	**Prediction indicators**
AAPL	workplace, shutter, music
AMZN	healthcare, HBO, cloud
FB	unfriended, troll, politician
GOOG	advertisers, artificial intelligence, shake-up
HPQ	China, inventions, Pittsburg
IBM	64 GB, redesign, outage
MSFT	interactive, security, Broadcom
ORCL	corporate, investing, multimedia
TSLA	prices, testers, controversy
YHOO	entertainment, leadership, investment

of PCG's Lasso Granger method to produce sparse graphs as compared to CGRAPH's Vector Auto Regressive model which used a fixed number (10) of incoming edges per node already pre-filtered by a semantic parser. This imposes an artificial bound on sparsity thereby losing valuable latent information. We overcome this in PCG using a suitable penalty term (λ) in the Lasso method.

Key PCG factors for 2013: The causal links in PCG are more generic (Table 5.2) than the ones described in CGRAPH, supporting the hypothesis that latent word relationships do exist that go beyond the scope of a single news article. The nodes of CGRAPH are tuples extracted from a semantic parser (SEMAFOR [78]) based on evidence of causality in a sentence. PCG poses no such restriction and derives topical (unfriended, FB) and inter-topical (healthcare, AMZN), sparse, latent and semantic relationships.

Inspecting the links and paths of PCG gives us qualitative insights into the context in which the word-word relationships were established. Since PCG is also capable of representing other stock time series as potential influencers in the network, we can use this to model the propagation

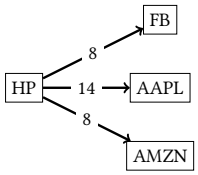


Figure 5.2: Inter-stock influencer links where one stock's movement indicates future movement of other stocks (time lag annotated edges)

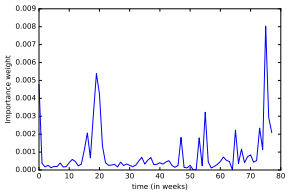
of shocks in the market as shown in Figure 5.2. However, these links were not used for prediction performance to maintain parity with our baseline.

5.5.3 Time Varying Causal Analysis

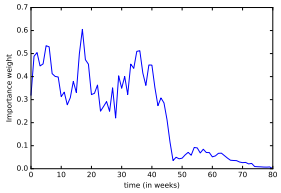
We quantitatively evaluate the time varying variant of PCG by using it to extract features for stock price prediction for a longer time window.

We present average root mean squared errors in Table 5.3 for different values of time windows of size W (50,100). For model selection, we used 50% of the time series and then used multiple time series of length W, disjoint from the ones used for time-varying factor selection and took average of the test error on the next 30 data points using the weights learnt. We repeat this using K-fold cross validation (K=10) for choosing the model selection data and present the average errors. The variation in importance weights of predictive causal factors for stock prices ("podcast", AAPL) and ("unfollowers", GOOG) is shown in Figure 5.3 which illustrates several peaks (for weeks) when the factor in the news was particularly predictive of the company's stock price and not significant during other weeks.

The time series and error graph shown for multiple stocks shows that the RMSE errors range between 1.5% 5% for all the test time series as shown in Figure 5.4. However, sudden spikes tend to display higher error rates due to the lack of training data which contain spikes. This issue is mitigated when the time window for training is increased. Increasing the window more than 100 did not improve the RMSE and came at the cost of training time. But, incorporating the



a “podcast” for predicting AAPL stock.



b “unfollowers” for predicting GOOG stock

Figure 5.3: Temporal variation in importance weight of predictive causal factors

spike residual PCG model, which predicts the leftover price value from the first model, provides significant improvements over the model without spike correction as seen in the last column on Table 5.3. Thus, we are able to achieve significant gains with an unsupervised approach without any domain specific feature engineering by estimating using an ARIMA model (p, d, q) = (30, 1, 10).

Table 5.3: Variation in stock price prediction error (RMSE) % with window size and spike correction

Stock	W=50	W=100	W=100 + spike
AABA	2.87	2.07	1.61
AAPL	2.95	2.84	2.28
AMZN	3.03	2.99	2.41
GOOG	2.67	2.36	1.91
HPQ	6.77	3.34	2.44
IBM	2.19	2.07	1.65
MSFT	3.03	9.45	4.80
ORCL	2.94	2.21	1.65
TSLA	5.56	5.52	4.32

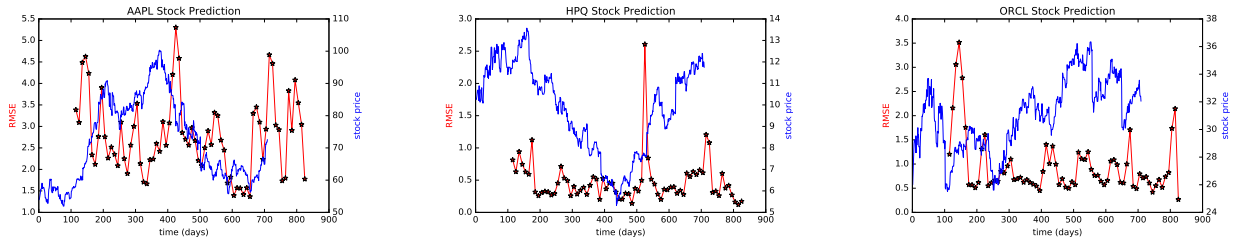


Figure 5.4: RMSE of stocks for longitudinal causal factor prediction without spike correction. Spikes in RMSE can be seen along with spikes in stock prices like HPQ.

5.6 Interpretation of Predictive Causal Factors

In order to qualitatively validate that the latent inter-topic edges learnt from the news stream is also humanly interpretable, we constructed PCG from the online archives of Times of India (TOI) [3], the most circulated Indian English newspaper. We used this dataset as, unlike the previous dataset which provided just the time series of words, we also have the raw text of the articles, which allowed us to perform manual causal signature verification. This dataset contains all the articles published in their online edition between January 1, 2006 and December 31, 2015 containing 1,538,932 articles.

5.6.1 Inter-topic edges of PCG

The influence network we constructed from the TOI dataset has 18,541 edges and 7,190 unigrams and bi-gram vertices. We were interested in the inter-topic non-associative relationships that PCG is expected to capture. We observe that a few topics (5) influence or are influenced by a large number of topics. Some of these highly influential topics are composed of words describing "Agriculture", "Politics", "Crime", etc. The ability of PCG to learn these edges between topical word clusters purely based on temporal predictive causal prediction further validates its use for design of extensive causal inference experiments.

[3] https://timesofindia.indiatimes.com/archive.cms

5.6.2 Causal evidence in PCG

To validate the causal links in PCG, we extracted 56 causation semantic frame [25] arguments which depict direct causal relationships in the news corpus. We narrowed down the search to words surrounding verbs which depict the notion of causality like "caused", "effect", "due to" and manually verified that these word pairs were indeed causal. We then searched the shortest path in PCG between these word pairs. For example, one of the news article mentioned that "Gujarat government has set aside a suggestion for *price hike* in electricity due to the Mundra Ultra Mega *Power Project*." and these corresponding causation arguments were captured by a direct link in PCG as shown in Table 5.4. 67% of the word pairs which were manually identified to be causal in the news text through causal indicator words such as "caused", were linked in PCG through direct edges, while the rest were linked through an intermediate relevant node. As seen in Table 5.4, the bi-gram involving the words and the intermediate words in the path provide the relevant context under which the causality is established. The time lags in the path show that the influence between events are at different time lags. We also qualitatively verified that two unrelated words are either not connected or have a path length greater than 2, which makes the relationship weak. The ability of PCG to validate such humanly understood causal pairs with temporal predictive causality can be used for better hypothesis testing.

5.7 Conclusion

We presented PCG, a framework for building predictive causal graphs which capture hidden relationships between words in text streams. PCG overcomes the limitations of contextual representation approaches and provides the framework to capture inter-document word and topical relationships spread across time to solve complex mining tasks from text streams. We demonstrated the power of these graphs in providing insights to answer causal hypotheses and extract-

Table 5.4: Comparison with manually identified influence from news articles

Pairs in news	Relevant paths in PCG
price, project	price-hike –(19)– power-project
land, budget	allot-land –(22)– railway-budget
price, land	price-hike –(12)– land
strike, law	terror-strike –(25)– law ministry
land, bill	land-reform –(25)– bill-pass
election, strike	election –(21)– Kerala government –(10)– strike
election, strike	election –(18)– Mumbai University –(14)– strike
election, strike	election –(20)– Shiromani Akali –(13)– strike

ing features to provide consistent, interpretable and accurate stock price prediction from news streams through honest estimation on unseen time series data.

6 | Granger-Causal Link Discovery in Large Temporal Networks

6.1 Introduction

Granger causality [156] in time series data is important in many real world applications in economics [16], climate science [258] and biology [257]. The knowledge of the Granger-causal structure allows us to build prediction models to make fine-grained time series forecasts conditioned on specific covariate values. For instance, the knowledge that a specific set of genes interfere with the expression of another gene allows us to build accurate gene regulatory networks, which assist in generating hypotheses for drug discovery. In practice, interventions are often infeasible because of the large dimensionality of data and making inference from real-time observations becomes inevitable as placing controls are either impractical, or even unethical. As a result, the inferences and forecasts must be done using only observational data. In this work, we assume that the underlying Granger-causal structure is specified to the extent that we know which covariates affect the outcome variable of interest. However, we lack information on how long the effect lasts and if it holds under all covariate distributions. For example, in the DREAM3 gene expression network task [342], where multiple genes can express and influence other genes, experiments are carried out where as part of a treatment, certain catalyzing agents are introduced over a time period and the corresponding gene expression time series are observed. Here, al-

though we know that there are a specific set of 100 genes that potentially Granger-cause each other, we do not know how the effect of "one gene regulating another" varies temporally. The number of time-lagged parameters in such a time series model can quickly grow with the maximum allowed time-lag (Table 6.1). In other words, because each of the gene expression varies over time, we need to know how the expression of one gene at one point in time regulates another gene's expression at a future time in a parameter efficient manner. This issue is unique to Granger-Causality over time series data in scenarios with limited data, relative to the high dimensionality of permissible time-lagged covariate distributions.

A common way to address this issue has been to be conservative and train prediction models over large time windows to capture any long-term effects of covariates. This approach often runs into data sparsity issues and poor granger-causal link discovery accuracy [390]. To deal with this issue, prior graphical granger methods [17] have artificially imposed sparsity constraints on the model parameters forcing co-efficients to collapse to zero, thereby reducing time-lagged parameters to estimate (from $600 - 700$ to ≤ 100). However, Granger-Causality was not designed for large temporal networks with millions of time-lagged parameters; in fact, economists have often warned against blindly applying it over a large number of variables [332]. One of the issues in applying Granger-causality directly to large temporal networks, is that large window sizes also result in the increase in chances of a violation of the positivity assumption [361, 413] - a condition necessary for consistency of the Graphical Granger methods, i.e certain time-lagged covariate distributions have been rarely or never observed previously and hence extrapolating the granger-causal links to such covariate distributions may be erroneous. In such scenarios, it might be best to defer prediction rather than predicting by extrapolating incorrectly.

To deal with the above challenges, in this paper, we propose a methodology to improve the recall of Granger-causal links by conditionally allowing for prediction deferrals. Specifically, given an outcome variable, a treatment variable, and a collection of covariates which are known to have passed the bivariate Granger-causal test (all of which are time series), our method parametrizes

each of the Granger-causal links with two parameters: (a) the maximum window size, δ, and (b) the variance threshold, ρ. The maximum window size specifies a bound on how long the treatment effect lasts and the variance threshold specifies when predictions are deferred. In particular, if the variance of our estimator for a given covariate value is above ρ, then we defer the prediction, suggesting that we do not have sufficient confidence in whether the treatment takes effect under the given covariate value based on the observations. Our method chooses the values of δ and ρ to optimize a Granger-causal link recall metric, which is computed using only those covariate distributions whose variance is below the threshold ρ. A small value of δ will result in fewer deferrals during training and thus a more reliable estimate of the recall metric, but might miss several links with longer time lags yielding lower recall, and the model becomes too sensitive to perturbations over smaller time windows. Very large values of δ also result in lower recall because they result in a large number of deferrals during training and consequently less evidence from data to support link discovery. Thus our formulation trades-off the model's temporal sensitivity and overlap-based robustness, and learn predictive models that have high accuracy and are consistent with the known Granger-causal links.

Prior work does not consider link recovery but instead focuses on optimizing prediction accuracy using general purpose sparsity inducing techniques. In particular, multivariate Auto-Regression linear models (VAR) are trained to optimize prediction accuracy while inducing sparsity in the time lag parameters through Group Lasso penalty [257] regularization; links are included by comparing the prediction accuracy of the model with and without the treatment variable [73] to test for significance. The non-linear version of the above VAR Granger Causality models have also been proposed [399] which could model additive effects of the past of each series in a decoupled manner. Sequence prediction models [454] and graph attention models [415] which model the neighborhood of nodes to learn Granger-causal links have also been studied. We build off of these constraints and demonstrate that augmenting the condition of overlap violation ensures that prediction models which learn to defer when specific covariate distributions have

not been previously observed are better at discovering Granger-causal links.

Learning such robust time-lagged Granger-causal models can be of immense importance in various real world scenarios of causal discovery. For example, in our running example of gene expression networks, using time series data from multiple such experiments carried out in laboratory settings along with our Granger-causal parametrization framework, we can extract optimal lag parameters between different genes to understand when (time-lag) and how (covariate overlap) they influence each other's expression. Similarly, in the human motion capture MoCAP task [69], we are able to improve the area under the precision-recall curve (AUC-PR) of detecting Granger-causal links in the human activity recognition dataset than baseline Granger-causality [154, 399] methods. Finally, given time series of Granger-causal news events, we improve monthly forecasting accuracy of stock prices. Each of these three tasks have a large set of time-lagged parameters (113K - 2.9M as per Table 6.1), but aim to predict a very small number of target variables. In such scenarios, the problem of over-parameterization may lead to prediction models that spuriously rely on multiple treatment time series variables. We overcome this limitation of prior sparsity inducing methods, by directly optimizing for the recall of Granger-causal links with sufficient covariate overlap.

The conditional temporal prediction models we have developed are applicable to a diverse set of forecasting tasks. Specifically, our conditional covariate based training approach has

- Reduced the number of parameters to learn by 2-3 orders, and achieved 25% better AUC-PR in discovering Granger-causal links than comparative baselines
- Improved prediction accuracy over held-out time series by 18-25% across three datasets in MoCap activity recognition, DREAM3 gene regulatory networks detection and the New York Times news-based stock price prediction tasks.
- Formalized the trade-off between time-lag sensitivity and overlap-based robustness and showed that artificially increasing the maximum time-lag leads to an over-specified model

with sub-optimal link recovery and prediction accuracy.

6.2 Granger Causal Link Recovery

6.2.1 Problem Setup

We consider a time-series forecasting problem where the goal is to predict the value $y(t+1)$ at time $t+1$ of an outcome times series using the past observations $\mathbf{X}(t,\delta) = \{\mathbf{x_1}(t,\delta), \mathbf{x_2}(t,\delta), ..\mathbf{x_n}(t,\delta)\}$, the treatment time series $\mathbf{v}(t,\delta)$, the historical outcomes $\mathbf{y}(t,\delta)$, each of them a vector of values evaluated up to δ discrete time steps back in time from timestamp t. We assume we are given a predictive model m, which outputs the future values of the outcome time series from the past observations:

$$\hat{y}(t+1) = m\left(\mathbf{X}(t,\delta), \mathbf{v}(t,\delta), \mathbf{y}(t,\delta)\right).$$

Each of the above variables $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}, \mathbf{v}, \mathbf{y}\}(t,\delta)$ are time-lagged multivariate variables with δ time-lagged values going back from time instant t. This means the predictive model has a total of $(n+2)\cdot\delta$ variables as input to predict the value of the outcome at time $t+1$. We also assume we are given non-parametric links of interest of the form $\mathbf{v} \rightarrow \mathbf{y}$. We now train this predictive model on historical observational data and use it to make fine-grained predictions for each covariate value $(\mathbf{x}, \mathbf{v})$. Implementing this approach requires us to answer two questions:

1. How long does the treatment effect last?

2. Under what covariate values does the treatment take effect?

Answering the first question allows to set the correct value of δ. The second question is important when dealing with observational data because the positivity assumption in covariate overlap $P(\mathbf{v}(t,\delta) = v|\mathbf{X}(t,\delta) = \mathbf{x}) > 0$ is often not met and blindly assuming it holds can lead to incorrect extrapolation.

Prior work has focused on answering the first question using sparsity inducing methods [17], while we argue that the second question is of equal importance to learn a robust time-lagged Granger-causal model and has implications on answering the first question. We formulate this as a joint learning problem with the *goal* to learn a compact Granger-causality aware time series predictive model which also learns to deferfrom making over-confident predictions for time periods with very few conditional treatment and covariate observations in the data [197].

6.2.2 Time-Lagged Granger-Causal Model Assumptions

Learning the time-lag parameters based on temporal predictions have been studied in the domain of Granger causality [154], where the links are established based on the lag parameter in a time series of the causal variable that provide the highest reduction in regression error of predicting the effect variable in a multivariate setting. These models make assumptions of sparsity, i.e for a given $(\mathbf{v}, \mathbf{y})$ only a small number of time-lagged variables of $\mathbf{X}, \mathbf{y}, \mathbf{v}$ are predictive of future values of $y(t+1)$. Many sparsity enforcing methods have been proposed like the Lasso regularization [17] which minimize the number of non-zero weights in a linear model [399, 26], or propose an auxiliary task based regularization of jointly predicting the causal graph and optimal predictors [235] or propose a recall-based regularization method to model autocorrelated time series with latent confounders [125]. One *overlooked assumption* in the above approaches in overcoming the overparameterization issue is that of the positivity assumption in covariate overlap ($P(\mathbf{v}(t,\delta) = v|\mathbf{X}(t,\delta) = \mathbf{x}) > 0$) [189], between treatment and outcome, given observational data. We overcome the limitation of prior methods by addressing violations of the positivity assumption explicitly through covariate conditional variance estimation.

6.2.3 Covariate Conditional Variance Estimation

Under the ignorability (no unobserved confounders) assumption, we can either estimate the importance of the treatment on the outcome variable from observed data by either slicing data based on treatment [375] or incorporating the treatment variable as another covariate [135, 236, 197]. In Algorithm 3, we adopt the latter approach and learn a time-lagged prediction model that optimizes the recall of Granger-causal links with the target variable y, using the covariates $\mathbf{X}$ and the treatment variable $\mathbf{v}$. Now we explain the Granger-causal link significance test used, and how we use a conditioned version of it to incorporate the covariate overlap assumption. Finally, we tie all of these components into a Bayesian optimization algorithm that maximizes the recall of the Granger-causal links.

Covariate Conditional Treatment Importance: To overcome the intractability of ensuring the overlap assumption for large number of covariates, we use the approach used by [197]. We estimate the lack of overlap using the conditional variance - $\hat{V}(\mathbf{x}, \delta)$ of the predictive models $\hat{m}, \tilde{m}$, which predict $\hat{y}(t+1)$ (Eqn 6.5). We use a non-parametric Conover Squared Ranks (SR) test-statistic used for testing for equality of variance in prediction errors [73], to approximate the *new information* that helps in improving the prediction accuracy of models. Once the prediction models are trained on certain splits of the data for a given outcome, we then estimate the variance by the bootstrapping method and evaluating the covariate variance on numerous held-out development splits. For a given threshold ρ, we adopt a trimming policy [180] with the rejection policy, conditioning on covariates $\mathbf{x}$, where the covariate variance $Var[\hat{V}(\mathbf{x}, \delta)]$ is above a threshold ρ [197], and then compute the *Time-Lagged Conditional Treatment Importance* - $TCTI(\delta, \rho)$.

$$TCTI(\delta, \rho) = \mathbb{E}_{\mathbf{x}: Var[\hat{V}(\mathbf{x}, \delta)] \leq \rho} \hat{V}(\mathbf{x}, \delta) \tag{6.1}$$

This formulation outlines that when the predictive causality test statistic has high variance,

we should not rely on those slices as they violate the overlap assumption. If the estimated value of TCTI passes the SR test with α significance ($\alpha = 0.05$) under a $F(1, (n+2)\delta + 1)$ distribution, we then consider the treatment variable v to have an α-significant Granger-causal link with y. This process is then repeated for all treatment variables to give the recall $w(\delta, \rho)$:

$$w(\delta, \rho) = \frac{\text{\# links } \alpha\text{-significant with } (\delta, \rho) \text{ as per Eqn 6.1}}{\text{\# Granger-causal links}} \tag{6.2}$$

For a given time-sensitivity lag parameter, the trade-off with overlap assumption to optimize recall can be understood by varying the hyper-parameters together. We implemented 3 approaches to fine-tune the hyper-parameters δ, ρ, including a grid search, random search and a Bayesian optimization technique [387] outlined in Algorithm 3, that directly models the recall of Granger-causal links as a utility function $w(\delta, \rho)$ (see detailed methods).

Algorithm 3 BayesOpt for Exploring the Time-Overlap Trade-offs

$(\delta, \rho) \sim$ U: uniform random distribution, $\hat{\delta} \leftarrow 0$
$M = \{\}, \hat{w} \leftarrow 0, \psi \leftarrow$stopping criterion
while $\hat{w} < 1 - \psi$ **do**
    Update BayesOpt acquisition function
    Acquire $D = \{k$ values of $\delta\}$ from BayesOpt
    Update $max_\delta \leftarrow max_{\delta \in D}(\delta, max_\delta)$
    Train $\hat{m}_{max_\delta}, \tilde{m}_{max_\delta}$ and update $\hat{\delta} \leftarrow max_\delta$ if $max_\delta > \hat{\delta}$
    Acquire $R = \{k$ values of $\rho\}$ from BayesOpt
    Update f, δ^*, ρ^* over R to maximize $w(\delta, \rho)$: Eqn (6.2)
end while
Return δ^*, ρ^*

Inference: Once we obtain the optimal δ^*, ρ^*, for each outcome variable, we use the trained models to make time-series predictions over unseen data. If the variance for a given covariate as pre-computed by the test statistic in Eqn 6.5 is above the threshold ρ^* at training time, we continue to defer to make predictions on those covariates at inference time. The resulting time-lagged prediction model then is used directly to infer the Granger-causal links based on the non-zero

model parameters in the models (Sec 8.1).

6.3 Results

We now demonstrate the efficiency of our approach in Granger-Causal link discovery, while showing that our prediction deferrals do not reduce prediction accuracy. Also, we show choosing the right set of temporal and covariate hyperparameters are critical for this improvement, as compared to generic sparsity inducing baselines where the choice is adhoc.

6.3.1 Granger-Causal Link Discovery

We demonstrate that our proposed method by optimizing $TCTI(\delta, \rho)$ and hence the Granger-causal link recall, we improve both the recall of Granger-causal links and the prediction accuracy across 3 datasets and 4 baseline Granger causal models (see baselines Sec 8.1 in detailed methods). In Figure 6.1, we see that the prediction accuracy (y-axis) improves by 18-21% and the recall of Granger-causal links (x-axis) improves by 25% across 3 datasets for each of the four prediction models' TCTI-{VAR, Neural Granger, Graph Generative, Graph Attention}, when the time lag and overlap parameters are optimized with 25 random restarts. The baseline models base-{VAR, Neural Granger, Graph Generative, Graph Attention} indicated as dots in the plot, have the sparsity constraint enforced through Lasso regularization loss, and end up with a lower recall of Granger-causal links i.e more Granger-causal links are not used for prediction, with low prediction accuracy. Further, to be comparable with baseline models which do not defer, we do not defer predictions at inference time, but only while learning the optimal time-lag and overlap parameters δ, ρ in Algorithm 3.

6.3.2 Importance of Prediction Deferrals

To understand how deferring predictions on covariates with high variance as per TCTI (δ, ρ) has helped learning better prediction models on the remaining covariates, we see that for each of the 3 tasks and 4 models in Figure 6.3, that there is an increase in prediction accuracy for the covariates we choose to predict as compared to the overall baseline model. Further, the increase is greater when the number of covariate slices deferred is greater. Thus by deferring predictions over covariates where overlap assumptions are violated, we improve the prediction accuracy and rely on robust Granger-causal links for prediction.

6.3.3 Variation in Time-Lag and Overlap Parameters

To understand how setting time-lag and overlap parameters is critical for the high performance of our approach, we plot the range of parameters required to achieve a fixed prediction accuracy across the 4 prediction models with our approach. In each of the dataset, we see that the distribution of parameters for time-lag, overlap-constraint for the links for a given Granger-causal model (the Graph Attention Model) has high variability as shown in Figure 6.4. Also, for a fixed prediction accuracy of outcomes, we see that the links for the 4 modeling choices are clustered into 3 groups - (low δ, high ρ), (low δ, low ρ), (high δ, low ρ). The lack of links with relatively high δ and high ρ further empirically affirms the trade-off between time-lag and overlap-constraints as shown in Figure 6.5.

6.3.4 Hyperparameter Optimization Method

We also see that across the 3 datasets, and the 4 modeling techniques - the choice of the hyperparameter fine-tuning methodology can impact the number of Granger-causal links we can recover with significant $TCTI(\delta, \rho)$ as shown in Figure 6.2. While grid search and random sampling provide good initial estimates of the maximum recall that can be achieved, we see that BayesOpt

quickly outperforms these brute force search mechanisms for the same number of model re-trainings. This drastically reduces the time and cost required to identify the temporal and overlap characteristics of the Granger-causal links.

6.4 Materials and Methods

6.4.1 Datasets

DREAM3 The DREAM3 gene expression network inference challenge [342] consists of 5 datasets, 2 for E.Coli and 3 for Yeast, with each dataset containing 100 time series. Each of the time series has 46 variables, each of them gene expression replicates observed at 21 time instants. For each of these 46 variables, we consider in a round-robin version that one of those variables is the outcome, one is the treatment, and the rest as covariates. This allows us to estimate $TCTI(\delta, \rho)$ for a total of 2070 combinations of treatment and outcomes, given the covariates. The ground truth contains directed Granger-causal links between 46 replicates, and through our prediction models, we parametrize each of the Granger-causal links by sweeping over values of δ, ρ with the maximum significant value of $TCTI(\delta, \rho)$. In parallel, we also report the AUROC (Area under the Receiver-Operator Curve) and AUC-PR (Area under the precision recall curve) for the classification task of detecting the binary gene expression.

MoCAP The CMU MoCAP dataset [69] consists of motion sensor data, for 54 joints, collected from two subjects for a total of 2024 time points. Here, we are given the Granger-causal links of the human skeleton and we learn the time lag and overlap parameters for the classification of each of the activities - jumping jacks, side twists, arm circles, etc. Here, we report the AUC-PR (Area under the Precision-Recall curve) for detecting the human activity based on the movements in the human joints, along with the recall of the Granger-causal links in the human skeleton.

Stock Price Prediction In the stock prediction task, the outcomes are each of the 10 stock

prices from 2010-13 (with 2013 as the test split), and for the treatment variable, we are given the top 10 financial news based Granger-causal factors and a further 100 covariates extracted from NY Times by [26]. Here too, we measure the Root Mean-Squared Errors (RMSE) of the predicted values and the fraction of time instants where we had to defer the prediction.

6.4.2 Methods

We now present the framework of our methodology to compute the trade-off parameters for the assumptions of temporal sensitivity and overlap-based robustness. By fine-tuning these parameters at development time to maximize recall of Granger-causal links, we learn the parameters that provide the best possible supporting evidence for the links in each of the domains. Note that sparsity inducing regularization constraints like Lasso, might lead to zero coefficient values for certain variables, and hence the recall of Granger-causal links, that measures whether all known Granger-causal links are used in the prediction model, may drop when optimizing for prediction accuracy.

Time-Lagged Predictive Causality

Under the ignorability (no unobserved confounders) assumption, we can either estimate the importance of the treatment on the outcome variable from observed data by either slicing data based on treatment [375] or incorporating the treatment variable as another covariate [135, 236, 197]. In this paper, we adopt the latter approach and learn a prediction model with high accuracy of the target variable y, using the covariates $\mathbf{X}$ and the treatment variable $\mathbf{v}$. To overcome the intractability of ensuring the overlap assumption for large number of covariates, we use the approach used by [197] to estimate the lack of overlap using the conditional variance of the predictive model m, which predicts $\hat{y}(t+1)$. (Section 8.1).

Since the values of importance weights can vary depending on the choice of models, for example in the case of linear regression - they are coefficients, whereas in non-linear network models, there are attention weights, activation vector alignment - similar to Granger methods [157], we

use a non-parametric Conover Squared Ranks (SR) test for equality of variance [73], to test if the treatment variable provides any *new information* that helps with the prediction accuracy of models: $\hat{m}$ with and $\tilde{m}$ without the treatment variable as input. The test is run on the prediction errors $\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1))$, $\epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1))$ produced by prediction models $\hat{m}$, $\tilde{m}$ respectively on held-out temporally disjoint test data.

$$\hat{y}_{(\mathbf{x},\delta)}(t+1) = \hat{m}(\mathbf{X}(t,\delta) = \mathbf{x}, \mathbf{v}(t,\delta), \mathbf{y}(t,\delta)) \tag{6.3}$$

$$\tilde{y}_{(\mathbf{x},\delta)}(t+1) = \tilde{m}(\mathbf{X}(t,\delta) = \mathbf{x}, \mathbf{y}(t,\delta)) \tag{6.4}$$

$$\hat{V}(\mathbf{x},\delta) = SR(\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1)), \epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1))) \tag{6.5}$$

The SR test we use, is the non-parametric alternative of the Levene's test [50], which itself is the robust alternative for non-normal distributions to the 1-way between-groups analysis of variance (ANOVA) [113] test to detect equality of population means. We use this test over the parametric ones as we do not make any assumption of the distribution of variance (normal), as our test of overlap violation cannot work if we already assume that there is an underlying normal distribution. The input for the prediction model are δ lagged time series of covariates, treatments and outcomes, and we will control the length of this time series as part of our methodology. We vary the δ and train jointly - warm starting hidden model parameters as the time lag δ increases, instead of training a separate model from random initialization per value of δ. Thus, we are able to compute the temporal lag that maximizes the prediction accuracy of the target variable that is Granger-causally linked to covariates. We characterize that a Granger-causal link to be supported by the observed data (or to be recalled), if adding a Granger-causal variable's temporal lag causes an increase in the prediction accuracy of the outcome conditioned on the covariate $\mathbf{X}(t,\delta) = \mathbf{x}$, as show by a test statistic with a p-value below the statistical significance threshold ($\alpha = 0.05$) under a $F(1, n+1)$ distribution, as compared to not incorporating the Granger-causal variable

at all. Otherwise, we characterize that Granger-causal link as not yet observed in the data. To convert the time-sensitivity into a variance based estimate, we compare the prediction errors ϵ and characterize it by the inequality of variance $\hat{V}(\mathbf{x}, \delta)$ given by the test statistic of the Squared Ranks (SR) test. Models $\hat{m}, \tilde{m}$ are trained and tested (Eqn 6.5) on temporally disjoint time series data.

Overlap-based Conditional Treatment Importance We now have to overcome the intractability of conditioning on the large number of covariates. Here too, we use the previously trained models: $\hat{m}, \tilde{m}$ to predict the target variable, but use the variance of the treatment importance estimate [197]. Once the prediction models $\hat{m}, \tilde{m}$ are trained on certain splits of the data for a given outcome, we can then estimate the variance by a bootstrapping method and evaluating $\hat{V}(\mathbf{x}, \delta)$ on numerous held-out development splits.

$$Var[\hat{V}(\mathbf{x}, \delta)] = Var[SR(\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1)), \epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1)))] \tag{6.6}$$

For a given threshold ρ, we adopt a trimming policy [180] with the rejection policy, conditioning on covariates where the variance of $Var[\hat{V}(\mathbf{x}, \delta)]$ is above a threshold ρ [197], and then compute the *Time-Lagged Conditional Treatment Importance* - $TCTI(\delta, \rho)$. While the models trained are dependent on δ, the computation of $TCTI(\delta, \rho)$ is done after the training is completed, rather than at training time.

This formulation clearly outlines that when the predictive causality test statistic has high variance, we should not rely on those slices as they violate the overlap assumption. For a given time-sensitivity lag parameter, this trade-off with overlap assumption can be understood by varying the hyper-parameters together. The utility function $w(\delta, \rho)$ is given by the fraction of the Granger-causal links in the given set that passes the difference in means t-test (with significance level α) that $TCTI$ is different as compared to the null distribution. We see that for thresholds: ρ,

such that the overlap condition $\mathbf{x} : Var[\hat{V}(\mathbf{x}, \delta)] \leq \rho$ is satisfied for all covariates $\mathbf{x}$, then there would be no deferral, and such a $TCTI(\delta, \rho)$ would directly evaluate the Granger-causality of the links, and hence $w(\delta, \rho)$ would be equal to 1.

We now outline the 3 approaches we undertake to fine-tune the hyper-parameters δ, ρ.

Grid Search: By using a grid search for values of $\delta \in \{1, 2, \ldots T\}$, $\rho \in \{\eta, 2\eta, \ldots k\eta\}$, for each Granger-causal link in the dataset, we search for the value δ^*, ρ^* that maximizes the $w(\delta, \rho)$. This way, we search among all Granger-causal links, the sensitivity and robustness parameters that is best supported by the observed data. This can be time-consuming to be done for each model and we can upper-bound the time lag parameter T to train the model.

Random Search: Instead of an exhaustive grid search, in this approach, we sample values of δ, ρ from a uniform distribution and choose the parameters that maximizes the $w(\delta, \rho)$. Here, we were able to heuristically choose a bound larger than $T, k\eta$ respectively and can search among values not explored by the grid search. Although we can control the number of hyper-parameters to train and evaluate the models against, the computational cost in training the models remain.

Bayesian Optimization: To learn which hyper-parameters δ, ρ result in high recall of Granger-causal links as per the significance level $\alpha = 0.05$, we used Bayesian optimization with the probability prior f parameterized by θ to be drawn from Gaussian processes. Specifically, we maximize the fraction of links w validated with a significance level for a given value of (δ, ρ). The covariance kernel chosen is the ARD Matern 5/2 Kernel [387], which has been demonstrated to capture realistic hyper-parameter distributions in neural networks, while resulting in sampling functions that are twice differentiable. We choose hyper-parameters in parallel using the Bayesian roll out method, where the acquisition function is optimized the utility of expected improvement per trial.

As noted in Section 6.4.2, δ requires a higher cost and time as it requires re-training of the model, while ρ can be fine-tuned post-training. These costs are modeled independent of the hyper-parameter distributions by calculating the expected inverse duration of computation and incorporating it in the expected improvement per second utility.

Inference: Once we obtain the optimal δ^*, ρ^*, for each outcome variable, we use the trained models $\hat{m}_{\delta^*}$ to make time-series predictions over unseen data. If the variance $\hat{V}(\mathbf{x}, \delta^*) > \rho^*$ for a given covariate $\mathbf{x}$ as pre-computed by the test statistic in Eqn 6.5 using $\tilde{m}_{\delta^*}$ at training time, we defer to make predictions on those covariates at inference time. Thus, we can now compare the predictions made by our overlap-aware Granger-causal model with baselines on the data slices where we do not defer.

6.4.3 Baselines

We would like to present a few comparable baseline prediction models built to incorporate Granger-causality for time series and discuss how compact time-lagged versions of these models have been learnt. These models form the baselines on which we evaluate in Section 6.3.

Multivariate linear auto-regressive models: Multivariate Auto-Regression linear models (VAR) take the time series of the treatment, covariates and the lagged values of the target variable as input to predict the target variable for future time instants. Here, we compare the prediction accuracy of the model with and without the treatment variable using the non-parametric Squared Rank test [73] to test the significance of a non-zero coefficient in matrix C. Additionally, we also add the Group Lasso penalty [257] that has been shown to overcome the need of precisely estimating the time lag by applying the Lasso regularization.

$$y(t+1) = A \cdot y(t, \delta) + B \cdot \mathbf{X}(t, \delta) + C \cdot \mathbf{v}(t, \delta) + D \tag{6.7}$$

Neural Granger causality: The non-linear version of the above VAR Granger Causality models have also been proposed [399] which could model additive effects of the past of each series in a decoupled manner. Here, by modeling the task of Granger causality using componentwise multilayer perceptrons and recurrent neural networks, all time series are captured in an input

layer of the neural network having a total of $\delta \cdot n^2 \cdot W$ parameters, where W is the number of hidden units in the input layer. In order to model long time lags in Granger causality, they use component-wise recurrent neural networks (RNN) for each time series. Similar to the linear model, to enforce sparsity, Lasso penalty and the hierarchical group Lasso penalty have been proposed, which chooses a suitable lag for each of the time series - but ignores the covariate overlap violation.

Generative Graph Neural Networks: Another approach proposed in [454], is to model this as a sequence prediction by reducing the graph to Breadth-First-Search (BFS) based deterministic sequence. They use a hierarchical graph RNN structure to first model the node prediction problem. In our case, although we know all the nodes of the network ahead of time, we can use the edge prediction model and predict edges in a BFS sequence. While this is comparable to the Neural Granger Causality model, the number of parameters to be learnt is lesser: $\delta \cdot n \cdot W$.

Graph Attention Model: Also recently, with the success of attention models in natural language tasks like machine translation, attending over the neighborhood of nodes, instead of recurrent architectures has been shown to be specifically relevant for graphical causal modeling [415]. This approach requires only the neighborhood of nodes and scales better than spectral representations of the graph, which need to be aware of the entire graph structure.

Other Related Work: We aim to understand the assumptions required to identify the optimal time sensitivity parameters of Granger-causal links in time series data, once the direction and presence of the Granger-causal links have already been defined. Prior work in this space has focused on methodologies to increase recall of the causal links in auto-correlated time series [137] or regularize over unseen parts of the causal graph [235]. However, such methods fail to quantify when it might be even possible to recover the optimal time lag parameters in an observed data distribution. The covariate relationship in time series have been explored in Granger causality with group boosting methods [257] or Markov random field regression [254] which capture the non-linear group information between variables in a time series. Prior methods [358] that main-

tain a set of probabilistic causal models and perform model selection can also benefit from the quantification of the trade-off between overlap and temporal parameters in longitudinal data.

6.5 Discussion

This paper provides a new framework for learning temporal and overlap parameters in Granger causal models for time series modeling tasks. These time-lagged models demonstrate significant gains compared to alternative formulations across three completely disparate time series tasks: news-based stock price prediction, the DREAM3 gene expression network analysis and MoCAP human motion recognition. The DREAM3 datasets and the associated inference research challenges have significant broader implications to the systems biology research community. Given our observed AUROC and AUC-PR gains, we hope that the systems biology community can benefit from our code base and model results (which we will release to the public). Similarly, the MoCAP dataset is widely used within the motion capture community. Here, we are able to demonstrate AUC-PR gains for detecting human activity and improve the recall on connecting these movements to Granger-causal links in the human skeletal structure. Our results on the stock market dataset would be highly relevant for the finance community and we believe this line of work can be extended to build causally-aware predictive models for socio-economic applications. Across all these open data sets, we believe our research conforms to the ethical guidelines outlined in these communities.

We have demonstrated the need to parameterize causal links with their associated temporal sensitivity with awareness of overlap assumption violations. In time series data, we show that there exists a trade-off between how temporally sensitive any prediction model incorporating the causal links can be while not compromising on the covariate overlap assumption. This allows us to further build better prediction models while not relying on data that lacks overlap while attempting to capture long term effects. Specifically, in the MoCap activity recognition task, we see

Dataset	n	δ	# Variables	# Targets
MoCAP	54	2,024	113,344	12
DREAM3	46	10,500	504,000	5
Stocks	100	29,200	2,978,400	10

Table 6.1: Problem of Over-parameterization in Time-Series Granger-Causal Models

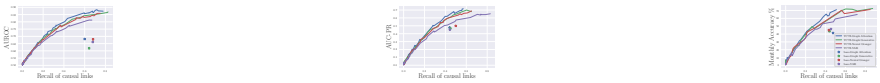


Figure 6.1: Recall of Granger-Causal Links vs Prediction Accuracy (a) DREAM3: Across each of the 5 outcomes in the DREAM3 dataset, we see that as the recall of the Granger-causal links in the gene expression network increases, the AUROC of the time series of the gene expression level also increases. **(b) MoCAP:** As the recall of the Granger-causal links of the human skeleton network increases, the AUC-PR in the human activity recognition task increases. **(c) Stock:** As the recall of the Granger-causal links of the financial news factors increases, the prediction accuracy of 10 stock prices increases.

that an inaccurate time-lag incorporated into the model can lead to inaccurate activity predicted which may have implications on applications in augmented and virtual reality. These errors emanate from incorrectly reconstructing the skeleton of the human body from the sensor data. In the DREAM3 gene regulatory network, if the expression of one gene if is incorrectly predicted, then we misinterpret one gene interacts with another - which may lead to ineffective drug candidates that target the gene regulatory networks. Finally, in the stock price prediction task, an incorrect time-lag can be catastrophic for any algorithmic trading application that relies on news based indicators - such crashes in the stock market have been anecdotally reported as flash crashes. Hence in all these scenarios, using the correct time-lagged model has implications in downstream applications, and if left unaddressed can lead to spurious Granger-causal links being incorporated. Further, since the size of such temporal networks are quite large with millions of parameters of estimate, a principled way of addressing covariate uncertainty through prediction deferrals can further improve the trust in practitioners.

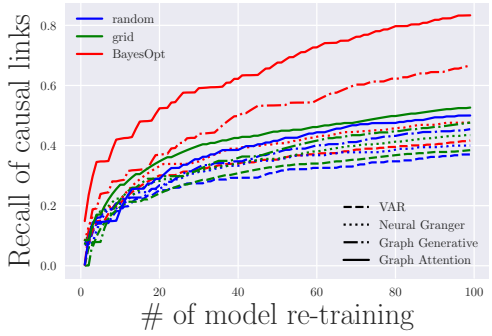


Figure 6.2: Sampling efficiency among hyperparameter search methods The number of re-trainings required to fine-tune the time sensitivity and robustness overlap parameters to improve the recall of Granger-causal links in the DREAM gene expression dataset

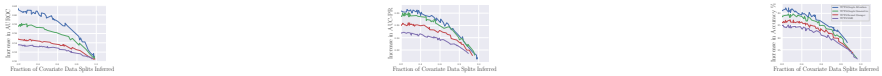


Figure 6.3: Prediction Deferrals effect on Accuracy Choosing δ based on an overlap based constraint, the prediction accuracy increases on the remaining test samples on (a) DREAM3 (b) MoCAP (c) Stock datasets for 4 Granger causal models. As the prediction models choose to defer on larger fractions of the covariate data splits, the increase in accuracy is higher.

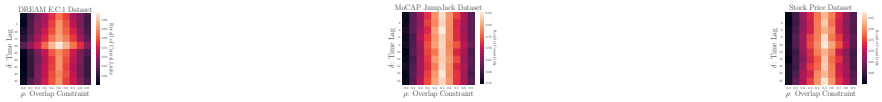


Figure 6.4: Variation between time lag and overlap parameters for the Graph Attention Model on 3 datasets shows the need to learn them jointly

Figure 6.5: Trade-off for a fixed prediction accuracy Time sensitivity and robustness overlap among the fine-tuned Neural Granger Causal prediction models across Granger-causal links that provide the highest $TCTI(\delta, \rho)$ for (a) DREAM3, (b) MoCAP and (c) Stock Price prediction datasets

Part III

Counterfactual Domain Reasoning

7 | Enhancing Domain Specific Concordance in Neural Recommenders

In this thesis, we focus on the design of **Domain Faithful Deep Learning Systems**, that translate expert-understandable domain knowledge and constraints to be faithfully incorporated into learning deep learning models. In high-stakes domains like health, socio-economic inference and content moderation, a fundamental roadblock for developing deep learning systems is that machine learning models' predictions diverge from established causal domain knowledge when deployed in the real world and fail to faithfully incorporate domain specific structure in counterfactual data distributions. Prior work in this space have formulated this problem as that of model generalization [298], data and label distribution change [251], domain adaptation [131], or adversarial robustness [70]. By doing so, they argue about model under-specification in the infinite data regime and data representativeness [75] over data distributions that are not realistically observed. While improving robustness of machine learning models is the core objective of all these approaches, they still fail to meet the expectations of domain experts on how machine learning models should behave when deployed in the real world.

Currently, domains where machine learning is being applied can be broadly distinguished based on the amount of prevalent enforceable domain knowledge in that domain. For example,

causal models [328] are robust and compact representations of domain knowledge which have implications of the conditional probabilities of the effect given the treatment and covariate distributions. Such an abstraction is common and well understood in industrial settings where the data generating procedure is well documented. Causal knowledge is often expressed in various forms - graphical causal models, semantic causal roles in sentences, theoretical model parameters. For example, causality based question answering lies at the core of customer support tools like chatbots. Prior ML models fail to capture the directed nature of causality, for example rain causes traffic delay, and not vice versa. Hence, learning asymmetric causal embeddings faithful to causal graphs can improve accuracy. Causal knowledge is also useful in data sparse conditions where interventions are often infeasible. For example, the task of forecasting famine is critical for the mobilization of aid to millions of people, but hard to solve due to data scarcity in fragile and poorer countries. By building a news-based causal-aware forecasting framework that extracts *causal features* from 11.2 million news articles across 2 decades in 15 fragile countries, we can improve forecasting accuracy compared to state-of-the-art predictive models.

On the other hand, even in domains where causal models are not established, certain counterfactual behavior of the machine learning models are expected. For example, trustworthy ML models in health recommendations need to be robust to medical concepts over unseen patient data, while traditional ML models focus only on optimizing accuracy over the observed but limited test data. By incorporating trust through doctor-specified mapping rules between diagnoses and medications through data augmentation, we can improve accuracy of state-of-the-art end-to-end neural models. Automated detection of online toxic comments improves the quality of interaction in social media. However, the variations in the context of comments make it hard to protect specific demographic groups from disparate impact. By explicitly modeling such nuances through *counterfactual data augmentation*, we can bridge the gap and improve the accuracy of detecting toxicity by 6%.

To overcome these limitations, I have developed domain faithful deep learning systems that

directly incorporate domain knowledge in various stages of the machine learning pipeline - model design, constrained optimization, data augmentation and feature selection. This has led to deployments of domain faithful ML systems for consequential socio-technical and natural language understanding tasks by collaborating with domain experts. Specifically, we address critical research questions such as "What data distributions do domain practitioners care about?", "How to faithfully convert domain knowledge into model constraints for better generalization?" and finally "How to evaluate whether the ML models we learn are grounded in the domain knowledge and in what ways do they deviate?". In doing so, we enable ML to be used towards positive socio-economic development, by tackling real-world societal problems in computational social science and NLP, and simultaneously addressing the fundamental ML research questions underlying these problems. Throughout this thesis, we adopt a research philosophy that strongly emphasizes "end-to-end system design", where algorithmic contributions are evaluated and deployed in the real world with the aim to adopt them at scale. For instance, the causal-aware and robust prediction models developed in collaboration with the World Bank and Google, have shown that relying on data alone can lead to incorporating spurious correlations, and low accuracy in data sparse or counterfactual scenarios, and hence, domain-specific structure is necessary for building robust predictive models. Overall, the research in the thesis has been focused on applying *domain faithful deep learning* to build causally faithful and heterogeneously robust predictive models in the domains of socio-economic inference, causal-aware deep learning, health, and toxicity detection. Each of these domains pose unique challenges on how to incorporate structure and the diverse techniques required to execute them. Now, we present the outline of the 4 sections of the thesis:

Domain Faithful Causal Models: Question Answering tasks power technologies like chatbots for customer support in businesses. Recent advances in machine learning for processing natural language text have broadly relied on large neural language models like Transformers which

capture the relationships between the word tokens in long sequences. The fine-tuning of these language models for multiple tasks have demonstrated state-of-the-art performance on benchmarks like GLUE. However, these fine-tuned models perform poorly on counterfactual sentences or inconsistently on downstream tasks which have specific structure like graphical causal models or domain-specific theory. In the causal-QA dataset [5], questions of the form "What causes X?" are posed, where X can be a disease, phenomenon and a real-world event. Neural Network models have been modified to predict causal links, but lack the consistency required, i.e undirected paths in a graph are still considered causal, whereas causal graphs are strictly directional. On the other hand, traditional Information Retrieval (IR) techniques that mine such causal information from knowledge graphs are limited in their generalizability to new and related terms mentioned in questions, i.e "flood" and "deluge" may have similar causes, but if "deluge" is not in the graph, then we have no way of estimating its cause. To overcome the limitations of using either an end-to-end model or domain knowledge as-is in its limited scale, we provide a way to incorporate the constraints imposed by the domain-specific structure - causal graphs in this case into BERT-like transformer based models. We demonstrate that when proximity between the embeddings of two nodes is modeled using a pseudo-quasi-metric, we are able to capture the directedness of causal graphs. Specifically, we measure three properties of *faithfulness* namely the uniformity of the embeddings, the correlation between distances of any two random nodes in the graph, and link prediction accuracy. In each of these graph-specific indicators, by imposing a regularization loss which penalizes inconsistencies in how the embeddings satisfy these two properties over two large causal graphs with 800K nodes, we obtain a fine-tuned embedding that not only achieves causal faithfulness better, but also improves the area under the Precision-Recall curve over the *Yahoo! Answers causal-QA dataset by 21%.*

Domain Faithful Feature Extraction: In socio-economic inference, the motivation is to have a broader positive societal impact using data-driven machine learning tools. Many applica-

tions which relied purely on data have faced issues as they did not incorporate domain-specific causal structure. For example, in the Flu prediction model based on Google Search Trends, it was shown that the model deviates over-time as compared to a one that incorporates signals derived from the Center for Disease Control (CDC). In the problem of predicting food insecurity task, we overcome the challenge of data sparsity in fragile states which are often encumbered with infrastructural and conflict-based issues that makes the task of data collection harder. As traditional indicators like rainfall, vegetation index, etc are often delayed, we aim to use the news streams published by reputed sources like BBC, Reuters, AP, etc. to automatically extract and construct causally grounded indicators. Our contributions extend beyond the methodologies and have implications on the ethical and operational trade-offs a domain practitioner needs to make in a socio-technical system. In the famine prediction task, by extracting causes from scientific literature using Semantic Frame Parsing and then constructing time-series indicators by expanding to tokens with low Word-Mover distances, we are able to *reduce the food insecurity forecasting errors by 32%*. Additionally, alignment of models to domain expertise provides an additional incentive to practitioners - counterfactual reasoning: Not all episodes of famine are the same, and our methodology allows us to model what is the implication of each of the causes in improving the prediction accuracy at a fine-grained level of districts in 15 of the most fragile countries in the world over two decades.

Domain Specific Concordance and Counterfactual Robustness: Recent advances in applying AI for healthcare have often relied purely on data, but fail categorically when patients with different characteristics than the ones present in training data are presented. Specifically, in the medication recommendation task, learning end-to-end neural models based on historical electronic health records might prove to be accurate, but may not inculcate trust in doctors, unless the ontologies of medicine that are used as standards by trusted medical associations are incorporated. In the medication recommendation task, since all possible diagnoses that may be

relevant might not be present in the training data, we improve the neural network model - G-BERT's *domain-specific concordance* based on expert-specified medical ontologies like medication and diagnostic code hierarchies and the mapping rules between them. By incorporating causal structure into machine learning models through categorical counterfactual data augmentation and regularization, we guard against predictions that violate the domain knowledge over categories and improve the *categorical robustness of prediction models by 1.2x* and accuracy by 12% on the MIMIC-III dataset, as we rely less on spurious correlations in the data.

Further, in the domain of toxicity detection in online social media comments, social-science experts have long advocated for incorporating how specific demographic groups are susceptible to specific types of toxic comments. It is important to model secondary attributes that are relevant to the toxicity of a sentence explicitly when we aim to be fair based on demographic groups. In this scenario, one needs to be aware of group-specific language, idioms, quirks, and background history to ascertain the toxicity of a comment. But this nuance was never captured explicitly in BERT-based neural network models. We incorporated this domain knowledge through counterfactual data augmentation that model secondary variables and were able to improve the ability to detect toxic comments for all demographic groups, specifically black women, who were susceptible to more directed toxic comments. By augmenting examples of directed toxicity in a weighted manner to demographic groups that are more exposed to such comments, we are able to classify toxicity better on all demographic groups. Without this nuance of how toxic comments vary, and just optimizing for overall absolute error, the toxicity detection model would disparately perform poorer on specific demographic groups unintentionally. Through intervention on secondary attributes through counterfactual data augmentation, we not only improved the model's understanding of what constitutes toxicity, but also improved the accuracy on all demographic groups by 7%. This application clearly demonstrates that as a text classification model is scaled to be applicable to all demographic groups in a society, the secondary effects of covariates and how they impact the performance of a ML system depends on domain knowledge, and needs carefully

expert supervision. Such business decisions and design choices have the capacity to influence the product experience for billions of users.

Domain Faithful Evaluation: Domain practitioners have often minimal guidance on the choice of parameters that AI tools in healthcare operate over. For example, in the angiographic disease status prediction task, the variability of diagnostic features in different demographic groups is well studied. Here, practitioners need to carefully evaluate the trade-offs between the per-group accuracy across demographic groups, when an end-to-end jointly trained model is used. When we analyze the performance of ML models on specific demographic groups, we outline the choice of parameters of fairness and accuracy trade-offs that practitioners have based on Pareto Efficiency. For example, how accurate an ML model should be over patients with darker skin tone than lighter skin tone in a heart disease status prediction model is a choice that cannot be made blindly, but with careful consideration of the medical diagnostic equipment's characteristics and the Pareto optimality of the model's performance across demographic groups. Through the principle of Pareto Efficiency, we can potentially *improve group-level accuracies by 9.6%* on UCI datasets. Acting blindly based on the neural model's decisions in high-stakes scenarios might be sub-optimal and using our methodology, experts can now justify their choice, in case they were to be contested.

7.1 Related Work

The notion of a model following a set of expert defined rules is prevalent in multiple domains of machine learning (ML) research. Below, we present a brief overview of these perspectives and how our approach aligns with them.

Hybrid Systems: Many approaches have been proposed to aid the domain expert in interpreting the machine learning model's predictions [136, 419]. Tools to guide the underlying deep

learning model through interactive feedback [53] and inductive logic [439] that increases diversity and aligns the model's predictions to expert knowledge have been proposed in the medical domain [300]. Applying data mining to extract association rules using Bayesian methods between input and output categories are also well studied [240], but they are typically not validated with rules by experts.

Interpretability: Mapping human interpretable rules with ML models has also been done to understand the inner workings of a black box machine learning model. For a broad review of the various notions of interpretability, we refer to [97]. Our work closely relates to the "task related latent dimensions of interpretability". Here, we care about the hypothesis of local interpretability[355], with incomplete coverage of domain expertise [462]. By restricting to this type of interpretability over expert-defined rules on subsets of the data, we seek that our models obey those rules.

Adversarial Robustness: To make machine learning models robust to perturbations, prior work has proposed defenses so that the model does not change it's output prediction for a small (ϵ), but humanly imperceptible change in the input [70, 54]. However, such adversarial robustness may either increase [188] or decrease [459] the overall accuracy of the models depending on the human specified notion of robustness. Hence, in the field of computer vision, robust models over concept based perturbations [445] and in natural language processing [185], robustness over word substitutions with synonyms are desired [344]. This indicates that the range of perturbations over which the robustness is defined, is equally important and going beyond geometrical definitions of robust boundaries is valuable [248, 346]. Hence, we choose to ground our models in expert defined relationships between inputs and outputs, which we would expect the non-observed data to generalize over.

Robustness in Recommenders: Recently, there has been a lot of interest in making recommender systems robust to avoid extremely undesired recommendations (e.g. horror films to children) [424, 448]. Robust models that explicitly guard against multiple attack models [187] like

profile injection [82], noisy ratings [311] and implicit issues like outliers [397], data not missing at random [230] have been proposed. Our definition is complementary to prior work in robust recommender models which propose simpler models like decision trees [227], fairness guarantees to avoid unintended bias [35, 83, 384, 40], temporal coherence to avoid catastrophic forgetting [424], defence against adversarial attacks of imperceptible changes [66, 174], and uncertainty based model calibration [448]. However, such approaches implicitly assume the presence of embeddings of items on which a similarity function (e.g cosine similarity) can be applied and assign a penalty if the recommender predicts items with low similarity. Instead, we explicitly use domain specific rules defined over categories of items and expect that the recommendations do not deviate categorically from those rules. Additionally, such approaches focus primarily on training-time attacks and do not address counterfactual scenarios that might arise during inference.

Substitutability: In recommender systems, the notion of substitutable items comes closest to the approach we take to create perturbations based on expert defined rules [281]. Such substitutable items have been inferred through browsing patterns like "users who viewed X bought Y" and co-purchasing logs [430]. Prior work incorporating categories through hierarchical autoencoders [93], multi-tasking [466], categorical embeddings [209] in recommender systems have improved accuracy. We combine these two insights and use expert provided rules to create category based substitutable counterfactual data to augment the existing training dataset.

7.2 Problem Formulation

We now present a formal description of our problem formulation and our goal to enhance neural recommender models through domain-specific concordance.

7.2.1 Notations

As illustrated in Figure **??**, in canonical recommender systems, each user has a discrete subset of historical items $X \subseteq \mathcal{X}$ (e.g., movies, diseases, etc.), which are then used to recommend to the user another subset of items $Y \subseteq \mathcal{Y}$, which may be of a different type (e.g., another movie, medicine). The recommendation problem is to train a model $h : \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{Y})$ given a dataset $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ ($\mathcal{P}$ denotes the power set). Our problem formulation works closely with the definition of categories of items that we can use to group recommended and historical items. This categorization based on individual item's characteristics is a choice in favor of discrete finite sets to describe the domain knowledge, and has been made in prior work [162] for easier reasoning by human experts. We assume the availability of such coarse-grained categories in our problem definition. Let's consider a finite number of discrete categories based on characteristics of the input items to be $j_1, j_2, ..j_n \in C_I$ (e.g., genres or part of the body). Each input item $x \in \mathcal{X}$ can be mapped to a subset of categories in C_I by applying the function $f_I : \mathcal{X} \rightarrow \mathcal{P}(C_I)$. Similarly, let's consider finite discrete output categories $k_1, k_2, .., k_m \in C_O$ and an output category set mapping function $f_O : \mathcal{Y} \rightarrow \mathcal{P}(C_O)$. We consider applications where there are priors between individual categories $j \in C_I$ and $k \in C_O$, that have been given by experts as domain knowledge. That is, we have knowledge of high level relationships between inputs and outputs that we expect the model to be mostly stable over. We represent these priors between individual categories using a mapping $p : C_I \rightarrow C_O$. This formalizes the expectation that for an input in a specific category $j \in C_I$, an output in a specific category $k \in C_O$ is recommended. We also consider that a distance metric d_c exists between any two categories, both over inputs: $d_c(j, j')$ and outputs: $d_c(k, k')$.

7.2.2 Medicine Domain Example

We illustrate the formulation of our problem with an example from the medical domain, where domain specific criteria are prevalent. In the MIMIC-III dataset, patient health data and their

corresponding visits to the hospital and medication are stored in electronic health records. The task of medication recommendation is to predict the set of medications prescribed by doctors by taking into account the patient's diagnostic codes, previous medication and other information. In this example, we consider the diagnostic ICD-9 codes (International Classification of Diseases) for a patient as input X. Each of the ICD-9 codes, $x \in X$ belong to an ontology of diagnostic codes, defined by a tree structure [318]. For example, consider the ICD-9 code "110.2" which describes "Dermatophytosis of hand", which belongs to the parent category "Dermatophytosis": j in the ICD tree. In our example, f_I is given by the *parent* function over the ICD-9 ontology tree. Also, let $y \in Y$ correspond to a recommended ATC (Anatomical Therapeutic Chemical Classification System) medication code [316], for example "J02AA" which describes "Antibiotics for systemic use". Similarly, f_O is the *parent* function in the ATC ontology which maps to the parent category k, which in our example is "Antimycotics". For the mapping between categories of diagnoses and medicine, there are expert-validated priors extracted from medical studies; for example in [**atc-icd**], the disease category "Dermatophytosis" j is mapped to the medicine category "Antimycotics" k. Each of these categories encapsulate a total of 10 ICD-9 codes and 3 ATC codes within them respectively. So, for instance, if the input ICD-9 code was: "Dermatophytosis of foot" (also in category j) instead of "Dermatophytosis of hand", then we, using the mappings from [**atc-icd**] as priors, we expect that one of the 3 medicines in category "Antimycotics" k would *likely* still be recommended.

7.2.3 Domain-Specific Concordance

Based on this understanding of examples and categories, we define now a set of perturbations and the concordance we expect over it.

Definition 7.1. Within-Category Perturbation: For an example $X \subseteq \mathcal{X}$ and a given input category j, we define a set $\delta_j(X)$ which contains perturbations of X by replacing a single item

$x \in X$ from category $j \in f_I(x)$ with another item also in category j:

$$\delta_j(X) = \{x' \cup X \setminus x | x \in X,\ x' \notin X,\ j \in f_I(x),\ \in f_I(x')\} \tag{7.1}$$

As defined, $\delta_j(X)$ offers a set of examples that, at least according to category j, are fairly similar to X. We now formally define concordance where such perturbations are done on a subset of the dataset $D_p \subseteq D$, which are covered by the domain-specific rules p.

Definition 7.2. Domain-Specific Concordance: For all examples $(X, Y) \in D_p \subseteq D$, such that $\exists x \in X, \exists y \in Y$ that matches a specified rule $p(j) = k$, i.e. $j \in f_I(x)$ and $k \in f_O(y)$, then we consider a model h to obey *domain-specific concordance* if for all within-category perturbations $X' \in \delta_j(X)$, we observe that $\exists y' \in h(X')$ such that $k \in f_O(y')$.

Stated more colloquially, whenever there is an example for which we see a relationship between the input and output that matches one of the domain expert rules p, we expect the model to be stable and continue to obey that rule over small changes that do not change the category of the input. Hence, we focus on changing one item at a time, and check if the outputs that had initially followed the category mapping continue to do so after the perturbation. This allows domain practitioners to reason about counterfactual changes in the inputs that do not modify input categories that are mapped by domain specific priors, and check for safe exploration within the boundaries specified by domain specific rules. However, we do not cover the scenarios when the input's categories do change, or when the example does not match an existing rule. Thus, we restrictively guard against sudden changes in a recommender model's output categories due to minor changes in the input whose categories remain unchanged. As motivated in the Introduction, in a movie recommender model, changing one "animation" movie to another in the user history, should not drastically change the category of all movies recommended from "animation" to say, "documentary". Specifically, we expect that at least one of the movies recommended still is an "animation" movie. Hence, our proposal is a hybrid framework where mappings between hu-

man interpretable categories can co-exist with neural recommender models. Having introduced the domain-specific category mappings, we now present recommender models that follow these category mappings.

7.3 Methods

Below, we present the methodology to optimize for robustness over the within-category perturbation dataset.

7.3.1 Rule-based Augmentation

In order to improve model robustness by reducing category misclassification, we define the category misclassification loss over within-category perturbations of examples in the observed dataset D as follows:

Definition 7.3. Category Misclassification Loss: For all examples $(X, Y) \in D_p \subseteq D$, such that $\exists x \in X, \exists y \in Y$ with $p(j) = k, j \in f_I(x) \wedge k \in f_O(y)$ and the indicator loss $\mathbb{I}$, the loss $\mathcal{L}_v$ due to misclassifying the output category k while the input changes from X to $X' \sim \delta_j(X)$ can be written as

$$\mathcal{L}_v(D_p) = \mathbb{E}_{(X,Y)\in D_p} \mathop{\mathbb{E}}_{\substack{(j,k):p(j)=k \\ X'\sim\delta_j(X)}} \mathbb{I}(k \notin \bigcup_{y'\in h(X')} f_O(y')) \tag{7.2}$$

We now have a loss over categories: L_v where we expect the output category to remain unchanged on counterfactual examples X' (Note that the above loss is non-differentiable and an approximation is provided in the following section). But, we still expect the exact label Y to be right for the original example X using the multi-label cross-entropy loss $\mathcal{L}$, measured using $\mathcal{L}_c$

as follows.

$$\mathcal{L}_c(D) = \mathbb{E}_{(X,Y)\in D}\mathcal{L}(h(X), Y) \tag{7.3}$$

Attempting to write a loss similar to (2), but on the actual counterfactual outputs Y', is difficult as we essentially do not observe them [329] and the changes are not imperceptible. However, by focusing on higher-level categories in (13.1), we expect that the categorical mapping p generalizes over unobserved counterfactual data (X', Y'). Expecting that models follow rules over categories of recommended items instead of specific counterfactual recommendations is what makes our framework easy to reason about, but also enforceable while training without having to *explain away* [437] all the counterfactual outputs by introducing more Bayesian priors. So, in order to improve robustness by training over Rule-based Augmented data (RA), while ensuring accuracy on the observational data, we combine the objectives using a α-weighted Lagrangian term to learn a new regularized model h_{RA}:

$$h_{RA} = argmin_h(\alpha\mathcal{L}_o(D_p) + (1-\alpha)\mathcal{L}_c(D)) \tag{7.4}$$

7.3.2 Within-Category Regularization

While h_{RA} minimizes the category misclassification loss over the rule-based augmented data, minimizing over all counterfactual perturbations $X' \in \delta_j(X)$ for a given rule $p(j) = k$ can be computationally expensive. However, minimizing the misclassification loss over a random sample of $\delta_j(X)$ can be less effective. To optimize for robustness in a principled sample efficient manner, we propose to regularize by minimizing, for each sample X', the upper bound of the difference between within-category output logits $z(X', y)$ and the observed output y logit which belonged to category k. By lowering this upper bound of difference between within-category logits and the observed output, we train the model to treat all items within a category as more likely than

items outside the category. We now formally define this Within-Category Regularization (WCR) loss.

Definition 7.4. Within-Category Regularization Loss: For an example $(X, Y) \in D_p$, following a rule $p(j) = k$, such that $\exists x \in X : j \in f_I(x)$ and $\exists y \in Y : k \in f_O(y)$ and $X' \in \delta_j(X)$; if $z_{(X,y)}$ denotes the logits of $h(X)$ for y, and $\mathcal{Y}_k = \{y' \in \mathcal{Y} | k \in f_O(y')\}$, then the *within-category regularization loss* is given by

$$\mathcal{L}_r(X, X', y) = max(0, max_{y' \in \mathcal{Y}_k}(z_{(X,y)} - z_{(X',y')})) \tag{7.5}$$

The expectation of $\mathcal{L}_r$ over all examples $(X, Y) \in D_p$ and all rules of the form $p(j) = k$ with X' sampled from $\delta_j(X)$ and y sampled from $Y \cap \mathcal{Y}_k$, give us the Rule-based Augmentation - Within-Category Regularization loss **(RA-WCR)**

$$\mathcal{L}_{ar}(D_p) = \mathop{\mathbb{E}}_{\substack{(X,Y) \in D_p, (j,k): p(j)=k \\ X' \sim \delta_j(X), y \sim Y \cap \mathcal{Y}_k}} \mathcal{L}_r(X, X', y) \tag{7.6}$$

Our approach is related to multiple lines of prior work. For example, interval bounded propagation [152] minimizes the upper bound of the output logits for inputs perturbed within ϵ distance in a l_∞ norm-bounded neighborhood. In our case, instead of perturbations defined in the l_∞ norm bounded neighborhood, we consider the set of within-category output classes. This also bares some similarity to the intuition behind distillation [183], logit pairing [210] and multi-task modeling [164] techniques. We adopt this technique as it smoothens the loss over a neighborhood of items within an output category instead of a strict cross-entropy category loss. A summary of the steps in RA-WCR is shown in Algorithm 5.

Algorithm 4 Rule-based Augmentation and Within-Category Regularization (RA-WCR)

Input: Dataset D, categories of recommended items (C_O) and input items C_I, and domain specific mapping $p : C_I \rightarrow C_O$
for all $(X, Y) \in D$ **do**
  if $(X, Y) \in D_p : p(j) = k$ **then**
    Sample perturbations $X' \sim \delta_j(X)$, $y \sim Y \cap \mathcal{Y}_k$
    Backpropagate $\alpha \mathcal{L}_{ar}$ over samples of (X', y)
  end if
  Back-propagate $(1 - \alpha)\mathcal{L}_c$
end for

7.3.3 Metrics

To build the neural recommender models that follow domain rules, we regularize the model such that within-category loss (13.2) is minimized. We evaluate improvement in robustness using the following distance metric between inputs.

Definition 7.5. Robustness Distance: Given all rules of the form $p(j) = k$, and the subset of the dataset D covered by them: D_p, *robustness distance* is measured as the average of the minimum categorical distance d_c between input categories j and j', where $x : j \in f_I(x)$ and a single item perturbation $x' \in S_k(X) : j' \in f_I(x')$ that leads to k being removed from the set of perturbed output categories $O(X')$.

$$O(X') = \{f_O(y') : \forall y' \in h(X')\} \tag{7.7}$$

$$S_k(X) = \{x' | X' = x' \cup X \setminus x \wedge x \in X \wedge k \notin O(X')\} \tag{7.8}$$

$$d_{robust} = \mathbb{E}_{(X,Y)\in D_p}[\min_{\substack{j\in f_I(x), j'\in f_I(x') \\ x\in X, p(j)=k, x'\in S_k(X)}} (d_c(j, j'))] \tag{7.9}$$

Using this, we can essentially answer the question, "Does the model follow the domain specific mapping between input and output categories?". For instance, consider the medical recommendation task where categorical distance d_c between inputs is defined as the distance between nodes of

the ICD-9 diagnostic ontology tree. Here, if the robustness distance $d_{robust} \geq 2$ for a recommender model, then we know that for the output category k to change, we need to perturb to an input x' in a different category, $j \notin f_I(x')$ (sibling nodes in a tree are at a distance of 2). Additionally, we continue to evaluate the change in the Jaccard similarity metric, F1 score and Precision-Recall Area under the curve (AUC) metric, Normalized Discounted Cumulative Gain on 100 relevant items (NDCG) [345] on the output classification task on the original held-out test data and also the *new* category classification task for the augmented within-category perturbation test data. In the next section, we will instantiate the categories: C_I, C_O, mappings: f_I, p, f_O for 3 domains of recommender systems. The ability to instantiate these finite category mappings based on the domain is one of the advantages of our hybrid framework.

7.4 Domain-Specific Instantiation

In this section, we will explain how the methodology described can be mapped to each of the three domains. *All examples are intended to test the usefulness of our framework, but the method should be adapted by practitioners and tested by domain experts for their needs.* As shown in Table 7.1, for the domains and rules we consider (Table 7.2), the rules do *not* suffer from low coverage ($|D_p| \ll |D|$) and can be used to augment and regularize.

Dataset	Total	Rules Applicable	Rules Violated
MIMIC-III	15,016	14,807	2,530
MovieLens	162,541	162,541	0
Last.fm	584,897	505,216	167

Table 7.1: Summary of total number of samples, samples where categorical rules are applicable and where they are violated in the observational datasets

For each of these domains, we define the current state-of-the-art model as ***Baseline***. As our framework incorporates more information through robust domain specific mappings through counterfactual augmented data, we also developed additional baselines that used these priors as input features. Specifically, we augmented categorical embeddings of each input to form the

Dataset	x	y	C_I	C_O	p	d_c
MIMIC-III	ICD	ATC	ICD-Tree Parent Nodes	ATC-Tree Parent Nodes	Expert-Defined	Tree Nod
MovieLens	Movie	Movie	Movie Tag	Movie Tag	Identity	Tag Scor
Last.fm	Song	Song	Genre, artist type, era	Genre, artist type, era	Identity	Hammin

Table 7.2: Instantiations of recommender systems into our hybrid framework

***Baseline*+Cat** model. In this baseline, no expert validation information is provided, but the category embedding is explicitly provided. We also augmented the embeddings of the applicable rule-based output category $k : p(j) = k$ as an input to the model to form the ***Baseline*+Mapped** model. This trains the model to pay attention to the mapped output category and minimize category misclassification. Finally, we instantiate our models ***Baseline* RA**, which modifies the baseline with Rule-based Augmentation (Eq. 13.1) and ***Baseline* RA-WCR**, which uses Rule-based Augmentation and Within-Category Regularization (Eq. 13.2). We set $\alpha = 0.2$ after cross-validation.

7.4.1 Medication Recommendation

We follow the MIMIC-III medication recommendation task as per [376], and the domain specific mappings p are obtained from [**atc-icd**] where medical experts validated a statistical table based on pairwise mutual information scores of co-occurrences between diagnostic x (ICD-9) and medication y (ATC) codes. These validated tables are segmented based on the age and gender of Austrian patients. Note that this dataset is different from the MIMIC-III dataset used in our evaluation. Hence, we use only the pairs of ICD-9: j, ATC categories: k that are expert validated p, but not any other statistical information from this study. A total of unique 349 pairs of ATC and ICD-9 Level 2 codes were deemed to be valid by the experts; 958 unique pairs if we break down by age and gender forms our domain specific mapping p. Age is bracketed into 3 ranges based on year of birth (1949-68, 1969-88, 1989-2008) and gender is considered to be binary (male, female). The categorical distance d_c used to define the robustness distance is given by the path distance between ICD-9 codes in the ICD-9 ontology tree. We use these validated pairs to generate perturbations in our existing dataset as shown in Algorithm 5.

7.4.1.1 Baseline

We use the current state-of-the-art for the medication recommendation task on MIMIC-III dataset as the *Baseline* - G-BERT [376]. This model uses graph embeddings based on the ontology of the ATC and ICD-9 codes. The model initially pre-trains the embeddings on the single-visit data using self-supervised learning, similar to BERT [85]. The graph embeddings are learnt using the Graph Attention technique [416], so as to learn hierarchical embeddings for each of the diagnostic and medication codes.

7.4.2 Movie Recommendation

In the MovieLens dataset [171], each movie x is tagged with user generated tags $j \in f_I(x)$, which illustrate different aspects like violence, thought-provoking, realistic, etc. We demonstrate the utility of our framework using an identity mapping $p(j) = k, j = k$ between movie tags in our analysis as shown in Algorithm 5. Colloquially, this means that if we see a user who has a history X of watching a specific category of movies, perturbing their history to a movie within the same category $X' \in \delta_j(X)$, should not completely drift the category of movies recommended away from that said category j. We measure categorical distance d_c using the absolute difference of movie tag relevance scores.

We would like to point out that the identity mapping p we have used is illustrative and more specific categorical rules could potentially help solve nuanced problems in recommendations, e.g., violent movies to children [173] or polarizing content with feedback loops [349]. To circumvent these pitfalls, lists of non-recommendable movies and simple human written rules are often applied. However, such rule-based post-processing approaches are often limited and there is an opportunity for these rules to be generalized over counterfactual data. Alternately, imposing rules on larger genres of movies like Romance, Crime is plausible using our methodology.

7.4.2.1 Baseline

As is common in MovieLens recommendation tasks, we consider the movies where the user has given a star rating of 4 or 5 to be positives, while the rest are negative. In addition to the movie's id and category, we use the historical ratings provided by the user on movies and their categories to predict whether the given movie should be recommended or not (star rating of 4 or 5). We use the baseline that is currently high-ranking for the MovieLens recommendation task, Deep Interest Networks (DIN) [471].

7.4.3 Music Recommendation

The music recommendation task is taken up on the Million Song dataset from Last.fm [**msd**]. Here too, the task is to predict the recommendation scores of songs Y based on the user history X. For each of the 502,216 songs, genres and tags associated to them are publicly available in semantic ontology databases. We specifically cross reference the songs and artists in the Last.fm dataset with DBPedia [22] to extract the tuple of the artist's genre, song type and date of release as the category of the song j. Similar to the movie tag space, we generate perturbations in the songs that belong to the same song type, era (in decades) and artist's genre in each of the user history logs. We expect that such perturbations will not have an impact on the ⟨song type, era and genre of the artist⟩: k recommended as shown in Algorithm 5. Here too, the domain specific mapping p is an identity mapping. To evaluate the categorical distance d_c required to measure robustness, we use the hamming distance between the songs' tuples of ⟨song type, era, artist genre⟩.

7.4.3.1 Baseline

The baseline used is the current state-of-the-art, EASE, which uses shallow autoencoders [392] over the user history. By enforcing that the diagonal of the weight matrix to be zero, to avoid

	Model	Jaccard	F1	PR-AUC
Original	G-Bert	0.3679 ±0.01	0.5281 ±0.03	0.6212 ±0.03
	G-Bert+Cat	0.3564 ±0.02	0.5203 ±0.04	0.6146 ±0.03
	G-Bert+Mapped	0.3680 ±0.01	0.5299 ±0.03	0.6230 ±0.02
	G-Bert RA	0.3883 ±0.02	**0.5788** ±0.02	**0.6541** ±0.01
	G-Bert RA-WCR	**0.4300** ±0.01	**0.5967** ±0.01	**0.6775** ±0.02
Augmented	G-Bert	0.3677 ±0.03	0.5281 ±0.02	0.6199 ±0.00
	G-Bert+Cat	0.3301 ±0.03	0.5102 ±0.01	0.5952 ±0.01
	G-Bert+Mapped	0.3573 ±0.01	0.5249 ±0.02	0.6084 ±0.02
	G-Bert RA	0.3723 ±0.02	**0.5483** ±0.02	**0.6343** ±0.01
	G-Bert RA-WCR	**0.4033** ±0.01	**0.5699** ±0.02	**0.6596** ±0.02

Table 7.3: Our RA-WCR model improves accuracy metrics of G-BERT on the MIMIC-III medication recommendation task for the Original dataset and the category classification task for the within-category Augmented dataset

collapse to the trivial identity function, they learn the weights that capture the similarity between songs.

7.5 Evaluation

In this section, we evaluate our methodology on all three domains and five model structures from Section 7.4. For each domain, we study the impact of our method along multiple dimensions to confirm our hypothesis of whether it can improve accuracy (§13.8.1) and within-category concordance (§7.5.2). We further perform fine-grained evaluations to understand the source of the changes in accuracy and robustness by coverage, types of rules and popularity (§7.5.3). We use leave-one-out train/test splits for 10-fold cross-validation and report mean and standard deviation of accuracy and robustness, where the folds are generated based on equal partitioning of user IDs.

Model	AUC (original)	AUC (augmented)
DIN	0.7348 ±0.0034	0.7044 ±0.0021
DIN+Cat	0.7136 ±0.0017	0.6960 ±0.0076
DIN+Mapped	0.7236 ±0.0005	0.7057 ±0.0035
DIN RA	**0.7349** ±0.0002	0.7112 ±0.0025
DIN RA-WCR	**0.7351** ±0.0002	**0.7205** ±0.0028

Table 7.4: Our regularized version of DIN with Dice [471] improves the AUC for the movie recommendation task on the original MovieLens 20M dataset and the movie tag classification task on the augmented dataset)

Model	NDCG (original)	NDCG (augmented)
EASE	0.389 ±0.002	0.312 ±0.003
EASE+Cat	0.382 ±0.003	0.309 ±0.001
EASE+Mapped	0.389 ±0.002	0.312 ±0.003
EASE RA	0.389 ±0.001	**0.314** ±0.001
EASE RA-WCR	**0.394** ±0.002	**0.317** ±0.002

Table 7.5: Our regularized version of EASE for the Last.fm million song dataset improves the (Normalized Discounted Cumulative Gain) NDCG on 100 most relevant songs for both the original test data and the augmented test dataset.

7.5.1 Accuracy

To test if we improve accuracy on the original dataset, we evaluate overall accuracy metrics in Tables 3, 4 and 5. For the medication recommendation task as shown in Table 13.2, in the MIMIC-III diagnostic code classification task we *improve F1-score by 12.9%* with similar gains in Jaccard coefficient and PR-AUC and we *improve F1-score by 7.9%* on the medicine category classification task over the augmented dataset which contains counterfactual scenarios of in-category diagnostic codes, thereby increasing adherence to diagnostic-medication category mappings. As shown in Table 7.4, in the MovieLens dataset, we *improve AUC by 0.04%* in the movie recommendation task and *improve AUC by 2.2%* for movie tag classification on the augmented dataset. In the Last.fm dataset, we *improve NDCG@100 by 1.3% and 1.6%* on both the song and category classification tasks on the original and augmented datasets respectively. Across all three domains we observe clear improvements in accuracy not just on category classification for augmented data but also on recommendations in the original data distribution. Further, these improvements do

not merely come from making the category information available, but *how* they are used through rule-based augmentation. This suggests both that the domain specific rules are valuable and regularizing models for robustness aligned with these rules is an effective means to *generalize over both observational and counterfactual* scenarios.

7.5.2 Model Sensitivity

We now test: "Does our method effectively increase adherence to the domain experts' mappings?" To measure if neural recommender models follow domain-specific rules, we evaluate the robustness distance as defined in *Definition 13.1*, limited to the subset of the data specified by the mappings. To continue the ICD-9 code based medication recommendation example, the changes would be quantified by the edge distance in the ICD-9 code ontology required to change the output ATC medication code. As shown in Table 13.3, our G-BERT RA-WCR model *achieves a robustness distance* $d_{robust} = 2.4 \geq 2$, suggesting that the model on average follows the expert-defined rules for counterfactuals near observed examples. Having a robustness distance greater than or equal to 2, *implies* that on average for any change in the recommended medication category, the model expects that the input diagnostic code category should have also changed.

In the MovieLens dataset (Table 13.3), this distance is quantified by the minimum change in the tag relevance score of the perturbed movie, before which the recommended movie has no relevance to the aforementioned tag. The relevance scores range from 0 to 1 and a higher robustness distance indicates invariance to changes within a movie tag (violence, drama, etc). Our model DIN RA-WCR *improves the robustness distance by* 2.1× as compared to the baseline DIN. It shows that on average, the relevance of a movie's tag in the user history has to decrease by 0.35 before we find that the recommended movie does not have that tag (relevance = 0). This indicates our model is less prone to spurious changes in recommendation tags with small changes in the movie's tag relevance.

In the Last.fm Million Songs dataset, the robustness distance is specified by the average of

Model Version	Baseline: G-BERT (MIMIC-III) d_{robust} (ICD-9 tree distance)	Baseline=DIN (MovieLens) d_{robust} (Tag Score Difference)	Baseline=EASE (Last.fm M d_{robust} (Hamming Distanc
Baseline	1.3 (1.0, 1.6)	0.11 (0.10, 0.12)	0.20 (0.12, 0.28)
Baseline+Cat	1.1 (1.0, 1.2)	0.13 (0.10, 0.16)	0.28 (0.23, 0.33)
Baseline+Mapped	1.2 (1.0, 1.4)	0.15 (0.11, 0.19)	0.31 (0.29, 0.33)
RA	1.7 (1.5, 1.9)	0.21 (0.18, 0.24)	0.42 (0.35, 0.49)
RA-WCR	**2.4 (2.1, 2.7)**	**0.35 (0.32, 0.38)**	**1.20 (1.11, 1.29)**

Table 7.6: Our method considerably increases the mean robustness distance (± standard deviation in brackets - see Def. 13.1) in medication, movie and song domains.

minimum Hamming distance between the tuples mentioning the era, song type and artist genre between observed songs and their within-category substitutes, for which there is a change in the output's tuple. Our model EASE (RA-WCR) *increases robustness distance to* $d_{robust} = 1.2$ which more significantly, crosses the threshold of 1. This implies that for a change in the recommended song's tuple of <era, song type, artist genre>, there needs to be on average one change ($d_{robust} > 1$) in the input tuple parameters, thus avoiding spurious output category changes.

7.5.3 Dissecting the gains

To understand where the gains in accuracy and robustness originate, we analyze slices of data and understand the source of the increase.

Coverage: In Figure 7.1, we slice the datasets into 2 subsets (D_p and $D \setminus D_p$) based on whether they are covered by the expert mappings or not. This separation is obtained in the medical dataset by augmenting data using 30% of the diagnostic code categories covered by the rules p. In the Movie and songs datasets too, we augmented the training dataset with counterfactual with-category perturbations on 30% of the categories and split the original test set into two subsets, one containing the augmented categories denoted as "covered" and the rest as "uncovered". We show in Figure 7.1 that the *improvement in accuracy of the covered subset is higher than the uncovered subset.* Still, for the uncovered subset, there is no degradation in accuracy. The change in accuracy and robustness is measured with respect to each of the unmodified state-of-the-art baselines. Further, *as the coverage of the rules increases, there is a corresponding increase in accu-*

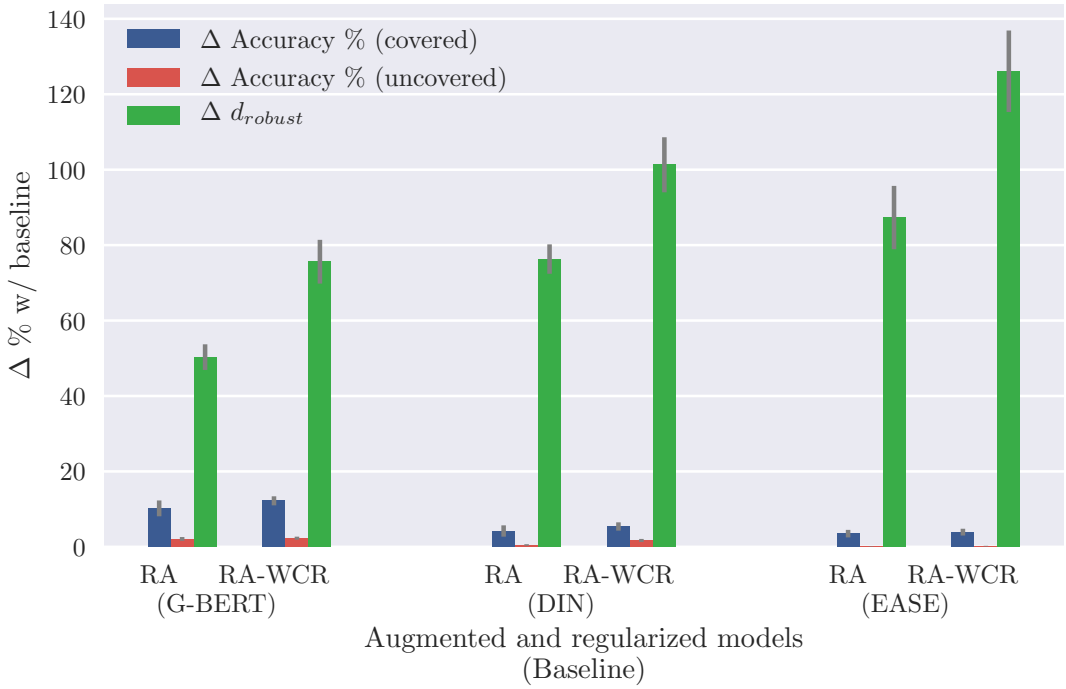


Figure 7.1: Our method improves robustness (bars are mean, with error bars showing one standard deviation) without degrading accuracy, and improves accuracy the most for subset of data covered by the domain specific mappings.

racy and robustness as shown in Figure 7.2. The numbers presented are averaged over 10 random samples of rules that cover a given coverage bracket for the medical recommendation task.

Domain Specific Rules vs Co-occurrence In this analysis, we explore which domain specific rules contribute to the highest gain in accuracy and robustness. This is to test our hypothesis that domain specific categorical rules that are not evident in the observed data are critical if we expect the model to generalize on counterfactual inputs. We bucketize the rules p based on a measure of co-occurrence: Normalized Mutual Information (NMI) score ρ between $(j,k) : j \in C_I, k \in C_O \wedge p(j) = k$ as observed in the dataset D. This allows us to differentiate between rules which are already supported by the observed data through sampling biases versus rules which are not. We bucketize the categorical mappings into five quintiles based on the NMI score in Figure 7.3, and show that *robustness gains obtained through rules which have low co-occurrence is higher than through rules which already have high co-occurrence in the observed*

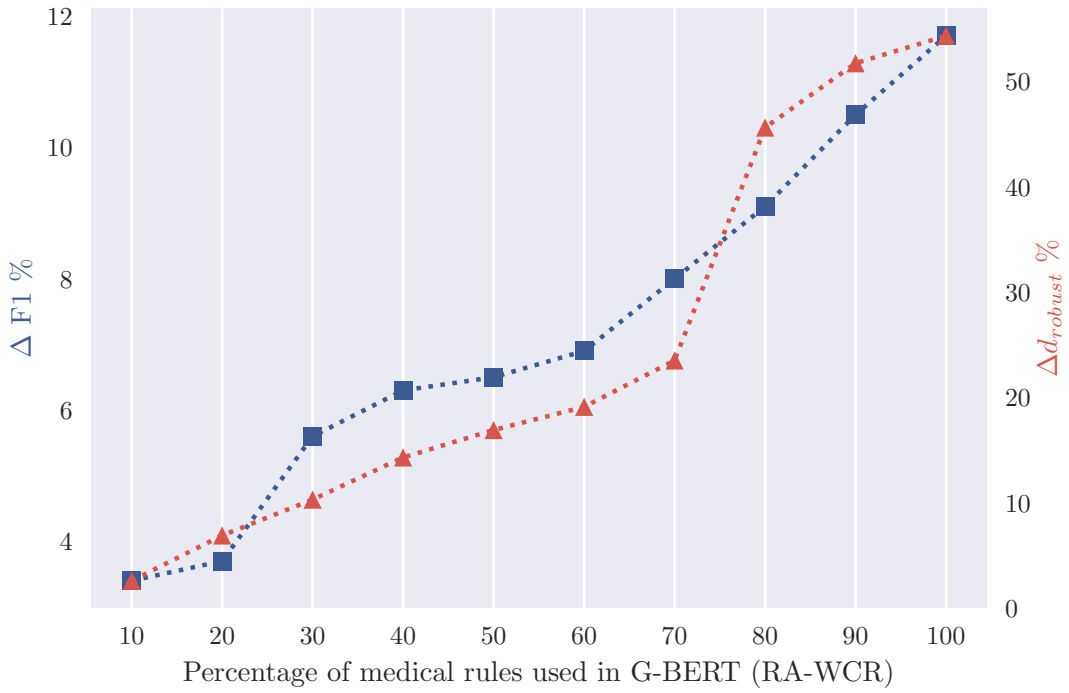


Figure 7.2: Our G-BERT (RA-WCR) model steadily improves F1 score and robustness distance as and when new medical rules are used to augment the dataset.

dataset.

Specifically, in the MIMIC-III dataset, we see significant gains in accuracy in addition to robustness when augmenting data using rules defined over medication and diagnosis categories with low NMI scores. This matches our hypothesis that there is value in obeying these expert-defined categorical rules. In the movie dataset, we bucketize based on the movie tag we augment the dataset by. We see in Figure 7.3, that augmenting data for rules based on *movie tags with high co-occurrence increases accuracy, whereas movie tags with low co-occurrence increases robustness on the original dataset.* This means that we improve robustness for niche movie tags with low co-occurrence like "sci-fi animation". A similar trend is observed with co-occurrence over music categorical rules in the Last.fm dataset.

Effect on Popular Items: In the MovieLens and Last.fm datasets, by following categorical rules, our robust models also tend to recommend popular items less frequently than the unmodi-

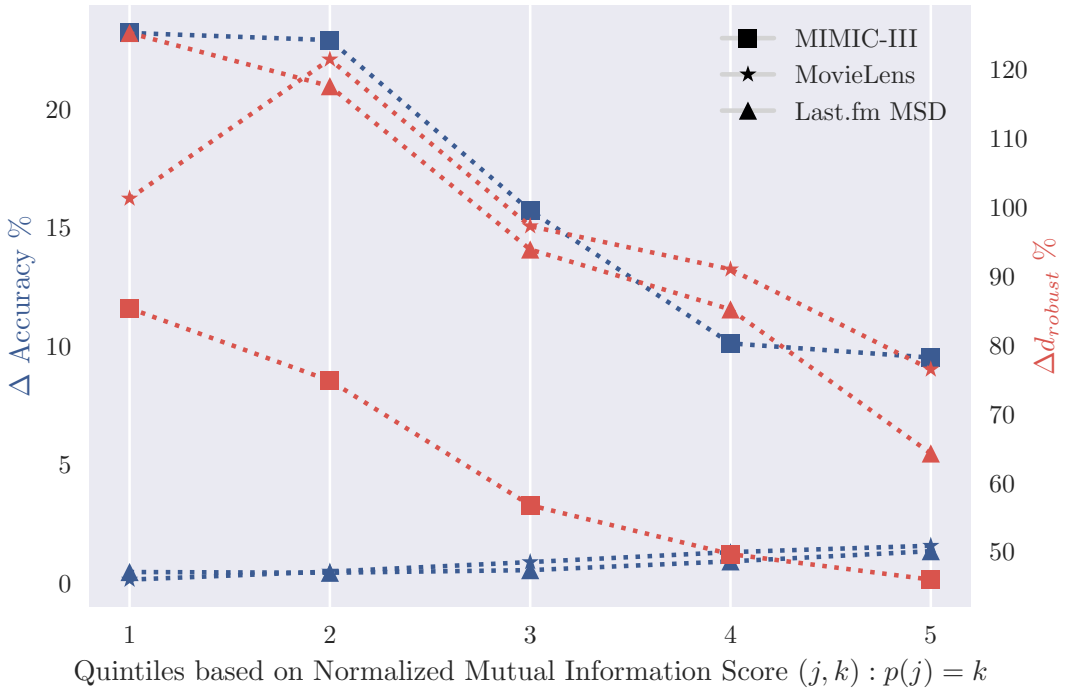


Figure 7.3: Our RA-WCR approach demonstrates more gain in sliced accuracy and robustness when augmentation is done through rules which have lower normalized mutual information score in the observed data across 3 domains

fied baselines, and rely more on the relevance to the tags than popularity in the observed dataset. In MovieLens, popular items (top-10 percentile) recommended *decreased by 32.3%* in DIN (RA-WCR) as compared to DIN. Similarly, in Last.fm, the number of times one of the songs from top-10 percentile were recommended *decreased by 23.8%* in EASE (RA-WCR) as compared to EASE.

7.6 Conclusion

In this paper we have laid out a novel framework for robustness and domain-specific concordance in recommender systems, based on within-category perturbations and expert-defined relations. We have proposed regularization based methods for using these expert-defined rules during training and demonstrated across three different domains that this improves not only the robustness of the recommenders, but also their accuracy. We believe this provides a solid foundation for further work in the community on how to enable domain experts to encode their expertise and define robustness based on that expertise in neural recommender models.

8 Improving Model Robustness through Secondary Attribute Counterfactuals

8.1 Introduction

How can we build NLP models that perform well over many slices, albeit sometimes small slices, of our data? Developing models that are *robust* in their performance is important for trusting these models to work well in diverse, unexpected settings. As a concrete running example in this paper, we will consider the task of toxicity detection: using a model to predict if a comment is toxic or not [95]. In this application, for example, it is often important to ensure that models are accurate over slices of data referring to different demographic groups, as has been raised across machine learning fairness research [170, 42].

One significant focus of research on how to improve model robustness has been addressing spurious correlations and improving *counterfactual* robustness. That is, researchers have found that models often rely on features or attributes that are only spuriously correlated with the task and accuracy often drops when models are evaluated on counterfactual data that perturbs those attributes [200]. To return to our example of toxicity detection, a model may learn that certain identity tokens are correlated with toxicity, but that could decrease accuracy for non-toxic

comments with those terms [95, 129]. Recent work has explored how counterfactual generation techniques can be used to form general checklists to *test* for model biases [357, 31], often composing many sub-problems which are hard to solve formally. Similarly, a wide breadth of research has studied how to *train* models to be more robust. We focus on one such mitigation technique—counterfactual data augmentation (CDA), where the supervised training data is augmented and balanced by replacing in-place words or phrases in the input sentence, which should not lead to a change in the output label Y [259, 475]. These counterfactual data generation approaches have been built on, as well as coupled with regularization, to improve counterfactual fairness [233], such as preventing models from being overly sensitive to identity terms [129, 337, 232, 326].

Although these approaches have been effective in reducing spurious correlations, in this paper we observe and study how such approaches often fail to significantly improve core model accuracy and can still perform worse on subsets of the dataset due to the primary variable over which counterfactual are generated being correlated with (many) secondary variables that are not swapped or balanced. Returning to our example task of toxicity classification over comments, the primary attributes (e.g., demographic identity terms) may be correlated with secondary attributes (intent of the comment—directed or descriptive) in the training data distribution. That is, for some demographic groups we may observe more directed comments and for others we may observe more directed comments:

Toxic: Seeking transgender rights is extreme (Directed) **Non-Toxic**: Transgender rights activists are labeled extremists (Descriptive)

In the toxicity classification example shown above, while the former is labeled as toxic by human annotators as it is *directed* towards a demographic group, the latter is only *describing* the toxicity and is considered as non-toxic. Nonetheless, both these sentences are classified as toxic by the Jigsaw Perspective API [95], thus leading to high false positive rates. So, to remove spurious correlations for the word "transgender" with toxicity, it may not be enough to improve model accuracy over comments with the word "transgender" if the model is more accurate for directed

comments than descriptive ones. Therefore we ask: *can explicitly considering counterfactuals over both primary and secondary attributes better improve robustness?*

To answer this question and improve model's robustness, i.e., accuracy on slices of the data, we propose a new approach, RDI, that learns from counterfactual data generated through interventions on *both* the primary and secondary attributes. RDI applies regularization techniques to train the model to disentangle the impact of the primary and secondary attribute and to explicitly optimize for the classifier's predictions to be sensitive or insensitive to each attribute. The approach also builds on recent reweighting approaches [219, 63] to further address distributional skews in the data.

Our approach to studying this problem builds on works that argue for a case-by-case analysis of variables and aims to provide a framework for incorporating secondary variables when we discuss the robustness of natural language models [141, 5]. Specifically, we have focused on the toxicity detection model which prior work has shown to suffer from unintended bias [95] based on protected identity terms mentioned in the sentence. We analyze how existing robustness techniques fail to capture a secondary attribute, namely the intent of the sentence while performing counterfactual data augmentation. We further show that this intent, that is descriptive or directed, is significantly correlated with specific protected identity groups in the dataset. By disentangling this correlation in the real world data via the counterfactual data, we obtain a model that does not disparately have high false positive rates on specific demographic groups, while being sensitive to the intent of the sentence. We achieve this improvement in robustness, while improving the sliced accuracy across multiple protected identity subgroups of the data.

Our **key contributions** are:

- We demonstrate how to disentangle the impact of protected and secondary attributes in NLP tasks like toxicity detection.
- We show how existing models perform poorly on counterfactual datasets that modify the secondary attributes, and train robust models that sensitize the model towards the sec-

ondary variables in a context-aware manner.

- Empirically, we demonstrate that our RDI method improves overall accuracy and sliced accuracy by 2-7% on all identity groups for both the toxicity detection task and generalizes on the coreference resolution task, while reducing spurious correlations through secondary attributes.

8.2 Related Work

Counterfactual Data Augmentation We build on prior work that performs counterfactual data augmentation [354, 44, 268]. Counterfactual data augmentation (CDA) has been used to create more balanced datasets to mitigate bias [260, 463, 474, 129] towards protected identity groups or improve accuracy [213]. Our work extends this literature by including a secondary variable that is correlated to the standard primary variable on which CDA is performed. This extension is motivated by works like [147] which demonstrate that there are secondary variables that need to be addressed for robustness.

Adversarial Robustness Making NLP models robust to adversarial perturbations has recently been explored extensively [472]. Work in this space define adversarial attacks through word or character perturbations [343, 104, 10] and certifiable defences [356, 201] following early work in adversarial training [148]. One of the challenges in applying adversarial techniques to the discrete domain of NLP is the lack of an ϵ-boundary in the input space. Hence, we consider only those interpretable perturbations that explicitly modify the primary and secondary attributes, as mentioned in a sentence.

Bias Mitigation Our work draws on recent works that aim to mitigate unintentional bias towards protected attributes in NLP tasks [45]. The approach of counterfactual token fairness which performs bias mitigation of template based [95] augmented data has been shown to improve

model performance over specific subgroups [129]. Debiasing techniques can be broadly categorized into in-processing: which changes training methodology [36, 203, 458] and post-processing: which operate post hoc on trained models [229]. While debiasing in unsupervised language models have also improved downstream tasks [433], we take the in-processing approach of debiasing in a supervised setting. Specifically, in the domain of coreference resolution, we closely relate to the work from [366, 112] to identify secondary variables; and in the domain of toxicity detection, we draw on qualitative error analysis [5, 115, 29] and domain expertise [431, 141, 368, 377] to derive our understanding of the secondary variable (intent of the comment) and how it relates to the label (toxicity); see Appendix 1. Another related perspective is that of distributional robustness where a machine learning model trained on one data distribution is evaluated on a modified data distribution [245, 263, 292, 252, 123, 15]. Following this body of work, our objective is to ensure that the model relies on invariances that generalize when the model is tested on slices of data, a type of distributional shift.

8.3 Problem Definition

8.3.1 Setup

Given a dataset, D, we will generate an augmented dataset, $\tilde{D}$ by adding synthetic, balanced and counterfactually augmented sentences.

Given an NLP classification task that operates on individual sentences $s \in D$, consider a primary variable X, which could be one of group based identities (say race, gender, etc) that is spuriously correlated with a secondary variable Z (e.g., intent of the comment—directed or descriptive) and the label Y that is to be predicted (say toxicity). In our setting, the values (x, z) of the primary and secondary variables X, Z are contained within an individual sentence s. We use the intent of the comment as our running example for Z in the toxicity detection task, but

our approach can be easily generalized to other factors like dialect, in-group language, figure of speech, etc. Note that since we are building prediction models that output $\hat{Y}$, we are interested in checking if a given model's predictions perform accurately on counterfactual inputs.

Our problem definition relies on the following assumptions about the primary and secondary variables prevalent in recent works on counterfactual robustness [200, 474, 219]. Firstly, given a sentence, the primary and secondary variables contained within it can be pre-specified. We also assume that counterfactual sentences that modify both the primary and secondary variables independently can be generated. Hence, we follow template based counterfactual data generation which specifies the primary and secondary variables in each sentence, as outlined in Section 8.5.2.

8.3.2 Objectives

Before we present our problem definition, we define the objectives that we will use from the robustness and fairness literature. Finally, we position these objectives within our context-aware counterfactual robustness problem formulation. For sake of simplicity here and in the following sections, we consider that the label, primary and secondary variables are binary with values $\{0, 1\}; \{x_0, x_1\}; \{z_0, z_1\}$ respectively. However, similar definitions for multivariate settings can be inferred.

8.3.2.1 Metrics

Original Dataset: In the original dataset D, as in most NLP tasks, we define the evaluation accuracy metric A over a set of sampled sentences s. Further, to evaluate the accuracy of the held out dataset conditional on the primary variable X, we compute the sliced accuracy $A(x)$ over that subset.

$$A = \mathbb{E}_{s \sim D} \mathbb{1}(\hat{Y}_s = Y_s) \tag{8.1}$$

$$A(x) = \mathbb{E}_{s \sim D} \mathbb{1}(\hat{Y}_s = Y_s | X = x) \tag{8.2}$$

Counterfactual Dataset: To improve counterfactual robustness, we aim to improve accuracy $\tilde{A}$ on the counterfactual dataset, by enumerating all possibilities of the values assigned to X and Z. We generate counterfactual sentences $t(s, x, z)$ by setting values of $X = x, Z = z$ in a sentence $s \in \tilde{D}$ using templates. Similar to overall accuracy, we can define sliced accuracy, $\tilde{A}(x)$ on the counterfactual dataset $\tilde{D}$ while enumerating all possible value assignments of the secondary variable. Note that the dataset $\tilde{D}$ represents a less biased dataset, one which might not actually be observed, but represents all possible values of the primary and secondary variables X, Z in $\tilde{D}$, and allows us to measure the toxicity detection model's counterfactual robustness around both the primary and secondary attributes.

$$\tilde{A} = \mathbb{E}_{\substack{s\sim\tilde{D}: \\ x\in\{x_0,x_1\}, \\ z\in\{z_0,z_1\}}} \mathbb{1}(\hat{Y}_{t(s,x,z)} = Y_{t(s,x,z)}) \tag{8.3}$$

$$\tilde{A}(x) = \mathbb{E}_{\substack{s\sim\tilde{D}: \\ z\in\{z_0,z_1\}}} \mathbb{1}(\hat{Y}_z = Y_z | X = x) \tag{8.4}$$

8.3.3 Goal:

Our robustness goal is to improve a model's robustness $A(x)$ - i.e accuracy on the original dataset sliced by the primary sensitive variable X. As secondary variables like Z are spuriously correlated with primary variables X in the original dataset D, we need to disentangle the impact of primary and secondary variables by optimizing on the generated counterfactual dataset $\tilde{D}$. In our paper, we achieve this goal by optimizing $\tilde{A}$, $\tilde{A}(x)$ over the dataset $\tilde{D}$, generated through interventions on both the primary and secondary variables, such that this improvement generalizes to the original dataset D.

8.4 Methodology

Since the goal of robustness is in addition to that of increasing overall accuracy on the original dataset, we use constrained optimization techniques over augmented counterfactual data. Before we present our proposed constraints, we present existing baseline constraints defined in the fairness and robustness literature. We discuss why these baseline constraints do not explicitly address the goal of improving counterfactual robustness on primary and secondary variables, and hence necessitate our additional proposed constraints on the counterfactual dataset $\tilde{D}$.

8.4.1 Baseline Constraints

Equality of Opportunity (EO): The Equality of Opportunity [169] constraint imposes statistical equality on the false positive errors, when conditioned on different values of the primary variable X. Such a constraint enforces that the primary variable X has no impact on the false positive rate of the model. We approximate this constraint over with the synthetic, balanced counterfactually augmented data $\tilde{D}$ (CDA) by minimizing the EO gap [464] with respect to the primary variable (Eqn 8.5) and denote it by the baseline "EO+CDA".

$$\begin{aligned} \min(|\, &\mathbb{E}_{s\sim\tilde{D}}(\hat{Y}_s = 1 | Y_s = 0, X = x_0) - \\ &\mathbb{E}_{s\sim\tilde{D}}(\hat{Y}_s = 1 | Y_s = 0, X = x_1)|) \end{aligned} \tag{8.5}$$

Counterfactual Token Fairness (CTF): In [129], the logits are equalized across counterfactual examples $s \sim \tilde{D}$ for different values of the primary variable X, but not the secondary variable Z. If $f(s)$ denotes the logit of the model's prediction, and $t(s, x)$ denotes the sentence generated by swapping the primary variable with x as per the template, then CTF minimizes the following

logit pairing gap:

$$\min \mathbb{E}_{s\sim\tilde{D}|X=x_0} |(f(s) - f(t(s, x_1))| \tag{8.6}$$

Since X and Z are spuriously correlated, both CTF and EO+CDA constrained models, which solely focus on X, are susceptible to performing poorly on examples when value of Z is altered explicitly. For example in the Jigsaw toxicity detection dataset, consider when Y is denoting "toxicity", X represents gender and Z the intent of the comment - descriptive or directed. If, for example, we observe in the real world that most directed comments are towards women, and not men (spurious correlation between X and Z), then just intervening on the gender X of the sentence and changing it from female to male, might unintentionally remove the impact of the secondary variable - the intent of the sentence, on the toxicity detection task Y. This is undesirable because the intent of the sentence is genuinely correlated with Y and its impact should not be removed.

8.4.2 Proposed Constraints

We overcome the limitation of not including secondary variable impact in baseline constraints, by explicitly modeling to *Maximize Secondary Sensitivity* in tasks like toxicity detection, where the label Y is sensitive to changing values of the secondary variable Z in the counterfactual dataset. We later discuss how this can be generalized to tasks where the secondary variable Z does not impact the label Y in Section 8.7.

Maximize Secondary Sensitivity: In some cases involving secondary variables, a characteristic that is often desired in a robust model is that it should be sensitive towards a change in a specific variable. For example in the Jigsaw toxicity (Y) dataset, even though more directed comments on online forums are towards females, and more descriptive comments are used for males,

the model should be sensitive to the intent of comment in determining the toxicity. If we blindly optimize for just CTF, the model may be less robust to changes in the intent of comments from descriptive to directed (Z). To overcome this issue, we propose a constraint that retains model sensitivity to changes in the secondary variable Z, while conditioning on the primary variable X. If $t'(s, x, z)$ is the template-generated sentence by swapping out values of x, z in a sentence s such that the label y assigned to the sentence changes to $\neg y$, and $f_y(s)$ denotes the logit of the model's prediction of y for s, then we propose to maximize the following conditional logit pairing gap.

$$\max \sum_{\substack{x \in \{x_0, x_1\} \\ y \in \{0,1\}}} \mathbb{E}_{\substack{s \sim \tilde{D} | Y_s = y, X = x, Z = z_0 \\ s' = t'(s,x,z_1)}} (f_y(s) - f_{\neg y}(s')) \tag{8.7}$$

Reweighting Samples All of the above constraints still do not enforce the independence between X and Z in the counterfactual dataset, $\tilde{D}$, if there is a sampling bias which prefers highly correlated samples of X, Z in D. This is because the real world dataset might suffer from selection bias, task annotator difficulty bias [151], etc, which cannot be easily offset through data augmentation alone. Therefore in addition to augmenting counterfactual data, we seek to reweight the augmented samples in such a way that the probability of Z conditional on X is equalized. Hence, a sentence $s \in \tilde{D}$ with $X = x_0, Z = z$ is weighted by w_s using an inverse-propensity based weighting [313] based on the prevalence of Z conditional on X. However, since we are fine-tuning over the counterfactual dataset to also generalize over the original dataset, we are concerned about improving residual accuracy. We, thus apply this weighting on only those samples in the original validation dataset which our unconstrained model has incorrectly predicted. This boosting inspired technique [371] emphasizes the need to equalize the prevalence conditioned on our worst-case examples [315] where our initial model $\hat{Y}_{base}$ has made an incorrect prediction. For example, we reweight based on the error rates, a sentence $s \in \tilde{D}$ with $X = x_0, Z = z, Y = y$.

$$w_s = \frac{P_D(Z=z|X=x_1, Y=y, \hat{Y}_{base}=\neg y)}{P_D(Z=z|X=x_0, Y=y, \hat{Y}_{base}=\neg y)} \tag{8.8}$$

Context-Aware Counterfactual Robustness Based on the relationship of the secondary variable with the label, we incorporate our proposed constraints on the counterfactually augmented dataset $\tilde{D}$ as a fine-tuning step. Thus, the methods we propose can be used on any NLP model as a fine-tuning task. We summarize our proposed RDI methodology based on the context of the secondary variable in Algorithm 1.

Algorithm 5 RDI (Reweight-Direct-Indirect)

1: Input: Trained NLP model -M's predictions $\hat{Y}_{base}$, primary variable X, secondary variable Z, label Y
2: **for** each batch **do**
3: Augment template based samples for all (X, Z) pairs to form $\tilde{D}$
4: Reweight samples based on (8.8)
5: $\mathcal{L} = \mathbb{E}_{s\sim\tilde{D}}\, CrossEnt(\hat{Y}_s, Y_s)$
6: $\mathcal{L}_{\mathcal{RDI}} \leftarrow$ (8.6) + (8.7)
7: Back-propagate $\alpha\mathcal{L} + (1-\alpha)\mathcal{L}_{\mathcal{RDI}}$ in M
8: **end for**

8.5 Evaluation

8.5.1 Data

The Jigsaw Kaggle toxicity dataset [1] contains sentences from the Civil Comment platform. We narrow down our focus to the comments that have the referenced identity in the comment, as well as the binary label: toxic or non-toxic. In total, 1,804,874 comments are annotated for toxicity, out of which ~50% of them have identities annotated too. Note that the identities are crowd sourced and not self-identified. We use a randomized 80-20 train-test split in our evaluation.

Some comments refer to certain protected identity groups, which we refer to as the primary

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

variable. Based on the qualitative study of toxic comments [431], we can broadly categorize the intent of comments as either directed or descriptive. Directed comments are speech towards a specific individual or group, whereas descriptive comments are more factual and do not hint towards a group or individual. Different identity groups are exposed to different intended comments, thus making the intent of the comment (descriptive or directed) our secondary variable Z. In this domain, our goal is to mitigate the impact of the primary variable on the prediction (Eqn 8.6), while retaining the sensitivity of the secondary variable on the predicted label (Eqn 8.7).

8.5.2 Augmentation Templates

The above dataset is the basis on which we evaluate the accuracy of the original dataset using our RDI algorithm. However, this dataset is not amenable for counterfactual data augmentation. Hence, we rely on a set of template based datasets to generate the counterfactual data on which we will fine-tune our models. [95] released a set of madlibs templates to generate toxic and non-toxic comments based on hierarchies of intersectional identities. We extend this framework to incorporate templates for intent of the comment - directed and descriptive based on the definition of toxicity provided in [431]. We provide an example of the 130,721 such counterfactual examples generated below (see appendix 1 for the full set of templates). Note that in addition to using templates, we can also utilize unsupervised learning based techniques to identify directed and descriptive comments.

8.5.3 Metrics

We evaluate the AUC for each identity group and the overall dataset in the Jigsaw Toxicity dataset. Since the secondary variable in the Toxicity dataset is not available for the Jigsaw dataset, we also present sliced AUCs based on the descriptive/directed intent of the comment as labeled by a model trained to predict solely the intent of comments with accuracy of 94.3% (details in Appendix 3).

Since we are comparing sliced accuracy across 9 identity groups in the toxicity dataset, we also compute the standard error bars in the measurement of each metric. We also perform a two sample independent t-test over $n = 10$ random restarts for each of the slices with $2n - 2$ degrees of freedom, and a significance threshold of $\frac{\alpha}{m}$, where $\alpha = 0.05$, $m = 9, 5$ (Bonferroni correction) for the two datasets respectively when we compare against the baselines.

8.5.4 Baselines

We present a brief description of the various baselines, each optimizing a baseline objective as discussed in Section 3.

Baseline	Model Objectives
Vanilla	Fine-tuned large uncased BERT model
EO+CDA	BERT+EO over balanced $\tilde{D}$ [464]
CTF+CDA	BERT+CTF controlled on the primary variable [129] over $\tilde{D}$
RDI	BERT + RDI algorithm

Table 8.1: Summarized description of baselines

Identity	Descriptive	Directed
black	0.58	1.00
white	0.77	1.00
gay	0.36	1.00
christian	1.00	1.00
jewish	1.00	1.00
muslim	1.00	1.00
male	0.97	0.99
female	0.97	1.00
blind	0.97	0.85

Figure 8.1: Accuracy of Jigsaw Perspective API model when sliced by the context (directed or descriptive) of the comments on our counterfactual dataset.

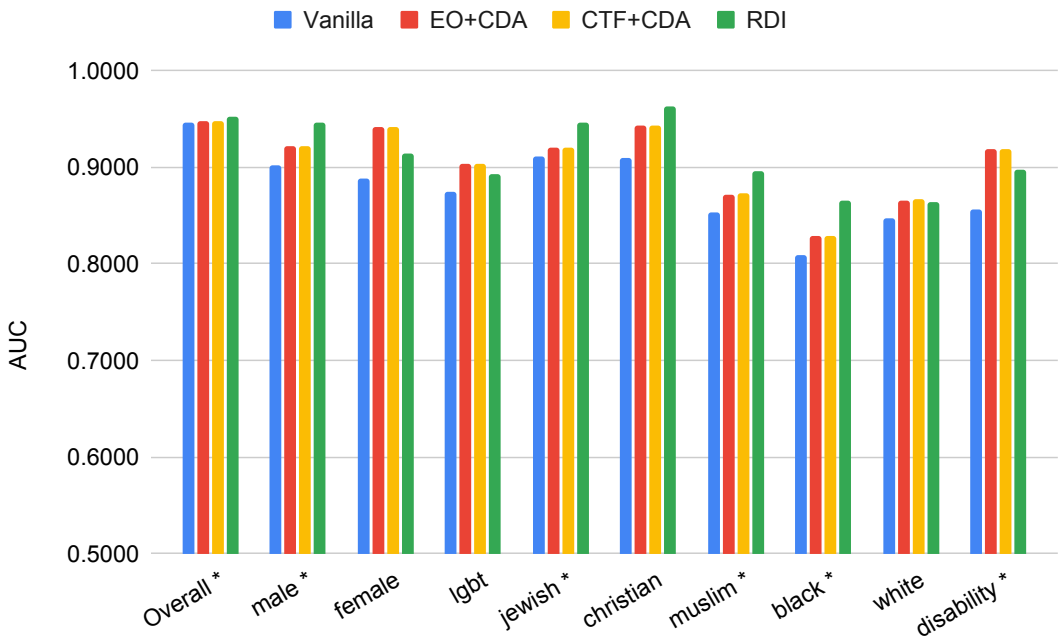


Figure 8.2: Area under the Curve (AUC) for toxicity detection across various demographic groups in the Jigsaw dataset

8.6 Results

8.6.1 Sliced Accuracy

In Figure 8.1, sliced accuracy of the vanilla model on the template based counterfactually augmented data highlights the need for improving sensitivity towards descriptive comments. In Figure 8.2, we show the impact on the AUC of identity groups as identified in the original Jigsaw toxicity dataset. Specifically, RDI performs 0.52% better in overall AUC with p-value = $0.001 \leq 0.005$ (significance level = $\frac{\alpha}{m}$), while increasing the **sliced AUC for black identity by 6.98%** (p-value=0.002). We see a general trend of improvement in AUC over the baseline vanilla model by 1.98–6.98%, with statistically significant improvement for groups of male, jewish, muslim, and black identities by making the model sensitive to the secondary variable – "comment intent". We subsequently fine-tuned a BERT model to predict the intent of the comment (de-

scriptive/directed) on the Kaggle dataset and sliced the change in accuracy as compared to the best performing CTF+CDA baseline. The resulting changes in Figure 8.3 demonstrate that for the slices where our model underperforms, it is due to a degradation in assessing directed comments for female, LGBT and disability groups. As expected, for the descriptive comments, we see consistent improvement across the board.

8.6.2 Ablation Studies

In order to understand the impact of the 3 objectives of the RDI algorithm, we conducted ablation studies by using the leave-one-out strategy (Figure 8.4). We note that, while removing the constraint based on counterfactual fairness (Eqn 8.6) has the highest impact, reweighting samples (Eqn 8.8) and controlling for secondary variables (Eqn 8.7) also have significant impact on both overall and sliced accuracy in the Jigsaw Toxicity evaluation dataset.

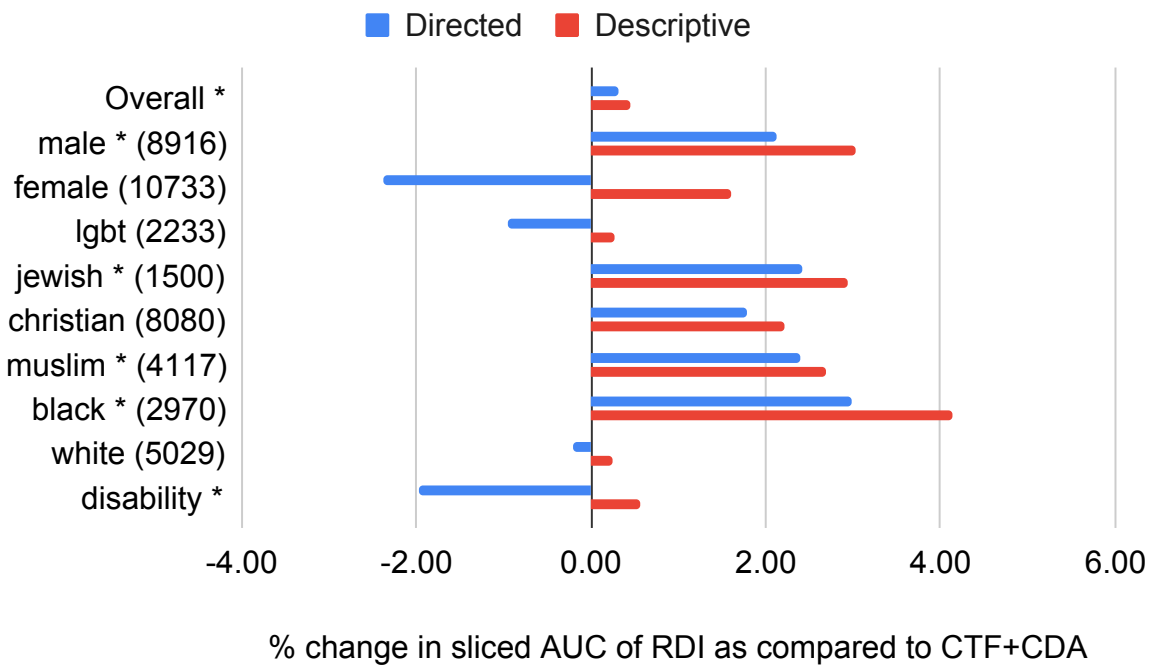


Figure 8.3: Change in Area under the Curve (AUC) for toxicity detection when sliced by the context (directed or descriptive) of the comments with slices with statistical significant change in asterisk.

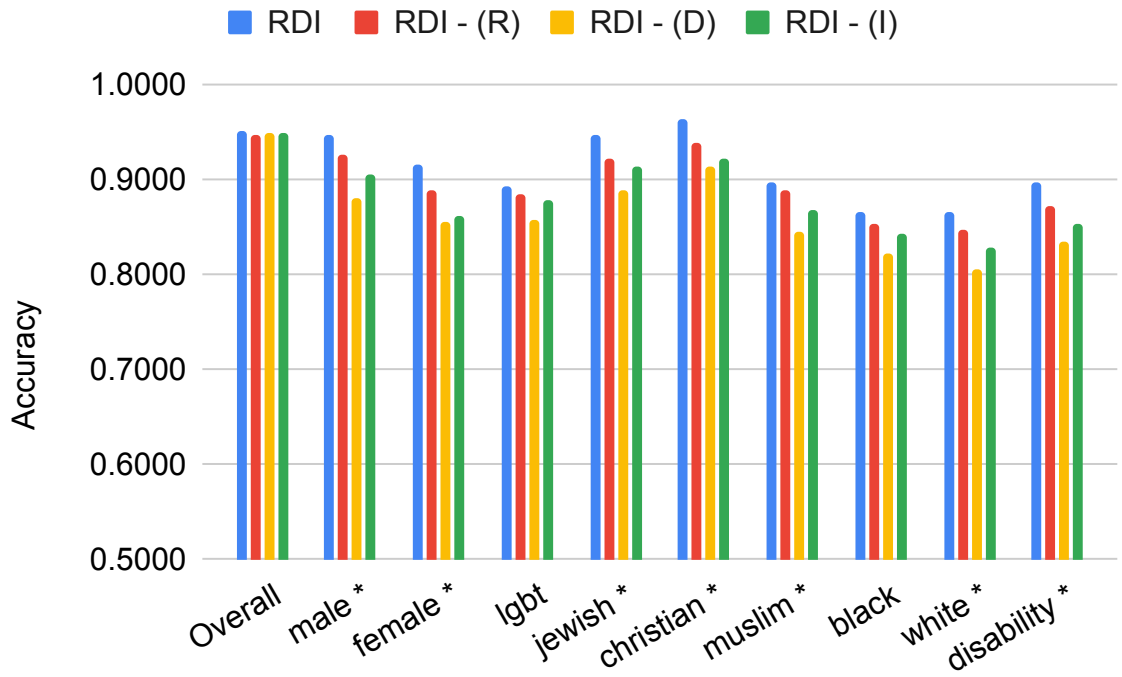


Figure 8.4: Ablation of the various objectives of RDI with slices having statistical significant denoted by *

8.6.3 Qualitative Analysis

We note that there is significant improvement in descriptive comments in most of the identity groups as shown in Figure 8.3. For example, in the black identity group, we see that the improvement in AUC is better in descriptive sentences 4.1% than directed ones 3.1%. While analyzing the errors of our model, we see that they occur often beyond the scope of our problem formulation [273] (Appendix 4).

8.7 Pronoun Coreference Resolution

We have demonstrated the utility of modeling secondary attributes to improve robustness of the toxicity detection models. However, we note that not all tasks have secondary attributes whose impact on the label needs to be maximized. Each task and their corresponding secondary

Metric	BERT-large-uncased	CDA	Dropout	CTF+CDA	RDI
F1-Score	0.93 ± 0.00	0.92 ± 0.01	0.88 ± 0.02∗	0.94 ± 0.01∗	**0.95** ±0.
Gendered Correlation	0.37 ± 0.03	0.25 ± 0.04∗	0.10 ± 0.02∗	0.23 ± 0.02∗	**0.11** ±0.
Gendered Profession Quintiles	Mean Gendered Pronoun Resolution % Female - % Male by Profession				
0-20	−25.2 ± 0.4	−23.8 ± 1.2	−17.1 ± 1.1∗	−21.2 ± 0.5∗	−12.7 ±
20-40	−18.5 ± 0.6	−12.8 ± 0.3∗	−9.1 ± 0.3∗	−14.5 ± 0.7∗	−8.8 ± 0
40-60	−11.5 ± 0.9	−10.5 ± 0.8	−8.0 ± 0.4∗	−12.7 ± 0.7	−5.9 ± 0
60-80	0.8 ± 0.4	0.5 ± 0.4	0.4 ± 0.5	1.6 ± 0.4	0.4 ± 0.
80-100	8.9 ± 0.2	7.0 ± 0.4∗	5.4 ± 0.5∗	9.3 ± 0.6	6.2 ± 0.

Table 8.2: Mitigating gendered correlation in coreference resolution as well increasing accuracy in the OntoNotes and Winogender datasets with statistical significant change denoted by *

attributes are unique in their relationship and their difficulty in data gathering, and we need careful understanding of the context while enforcing constraints between them. In this section, we show how our RDI framework can be extended to a task - "pronoun coreference resolution", where the label is invariant to the secondary attribute - gender. Between these two use cases, we have exhaustively covered the types of constraints that can be incorporated towards secondary attributes and encourage researchers to undertake a contextual treatment of secondary attributes in their tasks. We provide an example below, where the pronoun resolution should not change based on the gender of the pronoun.

Female: The nurse notified the patient that her shift would be ending in an hour. (her → nurse)

Male : The nurse notified the patient that his shift would be ending in an hour. (his → nurse)

Datasets and Augmentation Templates: For the pronoun coreference resolution task, we use the OntoNotes dataset shared as part of the CONLL 2011 and 2012 shared task [339, 338]. Each of the nouns referenced back from the pronouns also have their associated gender (binary) [32]. In the OntoNotes coreference dataset, we evaluate the F1-score, the gendered correlation coefficient [366] which measures the correlation between gender and the professions they resolve

to. The Winogender coreference resolution dataset provides templates with placeholders for the gendered-pronoun, and two antecedent professions which the pronoun could potentially be referencing. We refer to [366] for the full set of templates.

Label invariance to secondary variable In the gender bias Winograd dataset [366], the label is the coreference of the pronouns towards one of the two antecedents mentioned in the sentence. The pronouns are gendered (primary variable) binary - male and female; and the antecedents denote professions (secondary attribute) which the pronouns might get coreferenced to. Here, our goal is to minimize the unintended correlation of certain professions towards a specific gender. A systemic imbalance in the real world (see US Bureau of Labor Stats), is then reflected as a sampling bias in the text. For example, among the people with the profession "engineer", only 10.72% of them are females as per the labor statistics and a similar correlation is recorded in the text corpus, but an ethical ML practitioner would ideally want their robust model to **not** propagate these correlates by using the constraint in Eqn 8.9.

Minimize Secondary Impact: If we denote the logit of the model's prediction for a sentence s by $f(s)$, and the sentence generated by swapping out values of x, z in a sentence s without changing the label to be $t(s, x, z)$, then we propose to minimize the following conditional logit pairing gap, inspired by counterfactual indirect effects defined in [460] instead of Eqn 8.7.

$$\min \sum_{x \in \{x_0, x_1\}} \mathbb{E}_{\substack{s \sim \tilde{D} | Y_s = 0, X = x, Z = z_0 \\ s' = t(s,x,z_1)}} |f(s) - f(s')| + \mathbb{E}_{\substack{s \sim \tilde{D} | Y_s = 1, X = x, Z = z_0 \\ s' = t(s,x,z_1)}} |f(s) - f(s')| \tag{8.9}$$

Note that here, we explicitly focus on the change in error rates due to the change in the secondary variable Z, previously ignored by the baseline constraints. In the pronoun coreference resolution task, this amounts to equalized error rates on all professions, while conditioning on the gender of the pronoun X: male, female.

Robustness Gains: We see a similar trend in the overall accuracy for the coreference resolution task in Table 8.2. Here, we compare against one other baseline - dropout [433] where the baseline BERT model's dropout hyperparameters have been optimally finetuned for robustness. RDI outperforms existing baselines on both the F1-accuracy (higher is better) and the gendered correlation (lower is better). The lower gendered correlation also translates to a more even distribution of gendered pronoun resolution across the 5 quantiles of gendered professions [366].

8.8 Conclusion

We have demonstrated the value of incorporating the impact of secondary variables in the objectives for learning robust natural language processing models. We have shown that incorporating context-aware counterfactual robustness through the RDI algorithm, we improve performance on the counterfactual augmented data, but also improve the overall and sliced accuracy on the original dataset by 2–7%.

8.9 Broader Impact Statement

As we are dealing with the toxicity detection task, the concern of dual use for generating more toxic content on social media has to be considered. That being said, the identification of directed toxic comments towards minority communities can greatly improve the experience of members, often targeted due to their membership in protected classes in these online social communities. More so, when these same members describe the toxicity they experience on those social online forums, the possibility of them being flagged as toxic, can be harmful. We show that, without considering secondary variables, such errors, particularly in groups which are the target of toxic comments, can further exacerbate this divide. By developing an approach for controlling for known proxies, we hope this can enable practitioners to incorporate more domain knowledge,

particularly from users in under-served communities, to improve these systems. The template based counterfactual augmentation in capturing such nuances of secondary variables is a small step towards enabling more user participation and control in the design of these systems.

9 Improving Robustness through Pairwise Generative Counterfactual Data Augmentation

9.1 Introduction

Counterfactual data augmentation (CDA) has been used to make models robust to distribution shift and mitigate biases towards spuriously correlated attributes. Often, counterfactuals are generated as labeled examples through pre-specified templates [94, 166] or crowd-sourcing [214]. While natural text templates codify a specific number of assumptions of how counterfactual sentences and labels might vary, crowd-sourcing which can cover various types of counterfactuals, can be expensive. On the other hand, many existing methods [449, 468, 202, 10] simply rely on a *label-invariance* assumption: the label of the generated counterfactual example and the corresponding original example are the same. However, this simple label-invariance assumption does not always hold true [406, 306] and thus greatly increases the risk of using incorrect labels for counterfactual examples during training. For example, for many NLP tasks a small perturbation can easily change the ground-truth label [214, 128], e.g., changing the input from *This movie is great* to *This movie is supposed to be great* for sentiment classification, or changing the hypothesis from *The lady has three children* to *The lady has many children* for natural language inference.

Therefore, *"how can we automatically learn the labels for counterfactual examples, given a diverse counterfactual text generator?"* remains a challenging research problem.

Beyond costly human annotation or simplifying assumptions of label invariance, researchers have explored how to make use of a classifier f that has learnt to predict the label on the original dataset (X, Y). Such a classifier has been used to directly label generated examples (our "trust" baseline; [214]) or to weight generated examples based on the model uncertainty (our weighted-trust baseline; [323]). However, we see that using such simplistic labeling assumptions for counterfactual data augmentation have limited benefits for improving robustness, where we define robustness to be accuracy over a counterfactual test set of interest.

In this paper we propose an alternative approach to this problem: we leverage the sample efficiency of generative models to generate a large number of *diverse* counterfactuals, and train an *auxiliary classifier* which learn the *difference* between the original and counterfactual labels to annotate the generated counterfactual data. Specifically, we propose to learn the patterns of how counterfactual labels vary by using the *pair* of original and counterfactual sentences $(x, c_s(x))$ and the original label y as input to our pairwise classifier h and learn to predict the counterfactual label y'. The pipeline of our method is shown in Figure 9.1. We should note that only a very small set of human-annotated counterfactual examples are used to train the pairwise counterfactual classifier. Then in the inference stage, the pairwise counterfactual classifier is used to predict the labels for a large set of counterfactual examples. By using counterfactual generators and auxiliary pairwise counterfactual classifiers, we can greatly reduce the number of counterfactual examples for which we need human annotation, while providing similar gains in robustness comparable to a fully human annotated counterfactual dataset.

Our proposed approach addresses some of the challenges outlined in recent work like Checklists [357] in scaling the different types of counterfactual robustness desired in models beyond accuracy. We also show that each one of the counterfactual templates that the counterfactual generator produces contribute to a different type of robustness not previously captured by our

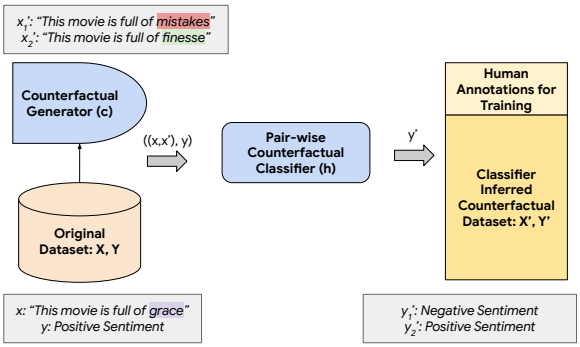


Figure 9.1: Overview of proposed approach: We propose a Pairwise Counterfactual Classifier to label generated counterfactuals (could be either label-invariant or label-modifying) at scale. We use the labeled counterfactuals as data augmentation and show it significantly improves robustness.

model, and hence further emphasizes the need to diversify the type of counterfactuals and generalize our performance against them for natural language classifiers. Thus, our paper provides a framework for incorporating diverse counterfactuals based on templates, by using generative models to scale the dataset and an auxiliary classifier to learn the label variance of the counterfactuals. Our core contributions in this work include:

- We propose a novel pairwise counterfactual classifier that labels counterfactually generated examples at scale based on a small set of annotated counterfactuals, improving *sample efficiency* of counterfactual data augmentation.
- We model both label-invariant and label-modifying counterfactuals for the sentiment classification task on Stanford Sentiment Treebank (SST-2) dataset, and the question paraphrase task on Quora Question Pair (QQP) dataset, and show robustness improvements using just 10% of human-annotated labels.
- The generated augmented dataset when used for fine-tuning produces an improvement in counterfactual robustness of 18-20%, comparable to a fully human annotated dataset, and a reduction in errors by 14-21% on IMDB, Amazon and Yelp reviews out-of-domain datasets that were not used during training.

9.2 Related Work

Our work is built on advances from various domains as outlined below:

Adversarial Text Generation Training against adversarial examples which perturb inputs in the vicinity of the existing training data by making geometric assumptions [306, 457] on a lower dimensionality of the data to improve robustness has been extensively studied recently. Natural examples which are syntactically and semantically similar to the original sentence, but produce different model predictions have been produced [10]. Similarly, defenses against adversarial attacks on self-attentive models have shown improvement in robustness to label invariant examples [185]. In FairGAN [449], they showed it is possible for a discriminator to achieve statistical parity on the real dataset, while performing the auxiliary task of detecting real and generated examples. Such controlled adversarial generative approaches [425] have demonstrated the effectiveness of automating data augmentation in text-based tasks. Generative models which optimize for fluency have passed human annotation checks where the model generated text is almost indistinguishable from human generated ones [265, 362]. We build on this body of work and utilize a generative model [442] that captures template-based counterfactuals to improve robustness. Generic adversarial notions of robustness however applicable, fail to incorporate specific counterfactuals directly in their training and orthogonal to our scope of study. Through carefully disentangling specific attributes and the rest of the latent variables in text, we generate counterfactuals across all possibilities, and utilize human-annotated templates to label a small fraction of the generated examples to train a pairwise counterfactual classifier.

Semi-Supervised and Self-Supervised Learning Labeling functions which provide crude estimates of the label have been used in semi-supervised methods [350], and are further used to learn a generative model to generalize over them. Further, utilizing unlabeled data [54] to improve adversarial robustness leverages geometric smoothing-based techniques to bridge the sample complexity gap between accuracy and robustness [451]. Thus, semi-supervised learn-

ing approaches aim to generate examples where the discriminator is least confident about [323]. Language models with very large number of parameters have also shown to be few-shot learners with minimal supervision [51]. Similarly, reinforcement learning based approaches with minimal labels have been proposed to combine the objectives of accuracy and counterfactual robustness [334]. Generalization against counterfactual examples by making models not to rely on salient features (easy examples) have been extensively studied by modeling biases in corpora [67, 68, 211, 299, 410, 139]. While the goal in these works have been building ensembles or end-to-end bias mitigation models, our goal is to minimize the number of human labels required to achieve an equivalent improvement in robustness. In this spirit of efficiently capturing the patterns already prevalent in the original dataset, and learning only the new ones introduced in the counterfactual templates, we learn the pairwise counterfactual classifier on a small number of samples, and use it to capture the label variations in the remaining counterfactual dataset.

Counterfactual Applications The counterfactual datasets we use throughout this paper were intended to highlight the shortcomings of existing models at the time. Improving robustness through training on the augmented data has been extensively explored [130, 443]. Learning how counterfactuals differ have been explored by comparing against gradient supervision [403] and the generalizability between original and counterfactuals [214]. The generated counterfactuals have also been used for explanations [418], highlighting biases [94] and debiasing through statistical methods [260]. This rich set of contrast sets [128], checklists [357], paraphrases [461, 438], adversarial schemes [367] and lexical diagnostic datasets [282] form the foundation of our method, which re-purposes them to build a counterfactual generative model and improve counterfactual robustness.

9.3 Methodology

9.3.1 Our Problem Framing

Let x, y be the input sentence and its associated label in the original dataset, respectively. We assume $y \in \{0, 1\}$ throughout the paper (i.e., we focus on binary classification tasks), but our framework can be extended to multi-class tasks as well.

Our core challenge is what is the true label y' for a generated counterfactual x'? Although we can further obtain human annotations, this can quickly become time consuming and budget intensive to do at scale. If we make the simplified assumption of label invariance throughout the counterfactual inputs x' generated, which is a common assumption in adversarial literature [149, 202, 10], we could end up with an incorrect counterfactual dataset which might hurt robustness and accuracy. Our goal is thus, to *generate a counterfactual augmentation dataset that produces a comparable improvement in accuracy and robustness as that of human-annotated counterfactuals with minimal supervision.*

We frame this problem as how to learn when the labels flip, i.e., identifying when the label of the counterfactual is different from the label of the original sentence: $P(y \neq y') = \delta$, $(0 < \delta < 1)$, in the counterfactual distribution $x' \in X'$. Given a generation model c, we denote $c_s(x)$ as the generated counterfactual over x by changing an attribute s in x. We also assume that a classifier $f : X \rightarrow Y$ has been learnt on the original dataset (X, Y) by optimizing for accuracy A. Since the counterfactual $c_s(x)$ can either contribute to a label flip or not, it is important for us to understand the patterns in the counterfactuals that vary the labels.

$$A = E_{(x,y)\in(X,Y)}\mathbb{I}(f(x) = y) \tag{9.1}$$

In our paper, the objective is to use the counterfactual data to train a model f' that improves

robustness, i.e., to make sure the models we trained generalize to unseen scenarios. We measure this by the counterfactual accuracy $\tilde{A}$ of f on a held-out counterfactual dataset (X', Y'):

$$\tilde{A} = E_{(x',y')\in(X',Y')}\mathbb{I}(f'(x') = y') \tag{9.2}$$

To achieve this goal, we generate our training counterfactual inputs $c_s(x) \in X'_t$ (here the subscript t denotes the training set) that modifies original input $x \in X$ based on the attribute s. In natural language tasks, the attribute s cannot be directly inferred from the sentence x and hence we rely on templates to define the types of counterfactual (e.g., negation, insertion, deletion) as commonly used in [357, 442] to infer the attribute s. Let $y \in Y, y' \in Y'_t$ be the label for the original and counterfactual sentences in our counterfactual training dataset. The training objective of robustness is to minimize the error $\mathcal{E}_t$ of the model f aggregated by attribute s on the training counterfactuals (X'_t, Y'_t), where CE refers to the cross-entropy loss, as follows:

$$\tilde{\mathcal{E}}_t(s) = E_{x\in X,(c_s(x),y')\in(X'_t,Y'_t)}CE(f(c_s(x)), y') \tag{9.3}$$

$$\tilde{\mathcal{E}}_t = E_{s\in S}\tilde{\mathcal{E}}_t(s) \tag{9.4}$$

Since y' is not readily available for counterfactual generated sentences $c_s(x)$ in our training dataset and gathering them for all examples can be expensive, our goal is to minimize the number of human-annotations of counterfactuals y' in the training dataset Y'_t, while achieving comparable improvement in robustness (Eqn 9.2). Hence, the training sentence and label set (X'_t, Y'_t) can be decomposed into two sets, one whose labels are human-annotated: (X'_a, Y'_a) and the other with model generated labels: (X'_g, Y'_g), such that $X'_t = X'_a \cup X'_g$, $Y'_t = Y'_a \cup Y'_g$. Our goal is to automatically learn the labels for counterfactual examples X'_g with an access to a limited human-annotated counterfactual data (X'_a, Y'_a), where $|Y'_a| \ll |Y'_g|$, while achievable counterfactual robustness $\tilde{A}$ (Eqn 9.2) comparable to the scenario when all the training labels are human-annotated.

9.3.2 Pairwise-Counterfactual (PC)

In order to generate labels for the counterfactuals, we construct a novel *auxiliary pairwise classifier* h, which takes in as input both the original dataset $(x, y) \in (X, Y)$, and a corresponding counterfactual $c_s(x) \in X'_t$ and the human-annotated labels $y' \in Y'_a$. The classifier h is trained on *pairs* of input sentences $x, c_s(x)$ and the original label y to predict $y' \in Y'_a$.

Specifically, the classifier h takes in the original input sentence x and its associated label y, as well as its corresponding counterfactual example $c_s(x)$. The output of the classifier $h(x, c_s(x), y)$ is the predicted label of the counterfactual example $c_s(x)$. In the training stage, the classifier h is optimized on the counterfactual examples with human-annotated labels $(c_s(x), y') \in (X'_a, Y'_a)$ via minimizing the loss function:

$$\ell_h = E_{\substack{(x,y)\in(X,Y) \\ (c_s(x),y')\in(X'_a,Y'_a)}} CE(h(x, c_s(x), y), y') \tag{9.5}$$

With the well-trained classifier h, we can generate the labels for any counterfactual example $c_s(x) \in X'_g$ (the counterfactual set without human annotation) according to:

$$y' = h(x, c_s(x), y) : (x, y) \in (X, Y), c_s(x) \in X'_g \tag{9.6}$$

9.3.3 Classifier-Aware Pairwise-Counterfactual (CAPC)

Additionally, since we know that f is already optimized to predict the label accurately on the original dataset, the auxiliary classifier h could potentially leverage f in its pairwise prediction through transfer learning. Specifically, if we decompose the counterfactual distribution (X', Y') as a mixture of samples from the original distribution (X, Y) and those that are independent of the original distribution, we would benefit by training h to identify samples from the latter distribution. In addition, assuming the correspondence between $f(x)$ and $f(c_s(x))$ is easier to

learn (e.g., with a lower model complexity), we could also benefit from learning a classifier-aware function to better capture this correspondence. Thus, we propose to augment the predictions of the original classifier $f(x)$, $f(c_s(x))$ as input to h as follows:

$$y' \in Y'_g = h(x, c_s(x), y, f(x), f(c_s(x))) : \\ (x, y) \in (X, Y), c_s(x) \in X'_g \tag{9.7}$$

Any uncertainty that f has on the counterfactual samples $P(f(c_s(x)) \neq y')$ can be mitigated by the auxiliary classifier h by identifying patterns in $c_s(x)$ when f predicts incorrectly. As a simple example, without any human annotation, the original model f might make incorrect assumptions on $c_s(x)$ that lead to incorrect predictions $f(c_s(x)) \neq y'$, e.g., a sentiment analysis model might give "positive" sentiment predictions due to the presence of qualifiers like "terrific", "amazing" (*this movie was amazing*) even when the counterfactual input $c_s(x)$ alters aspects of a sentence that changes the label (*this movie was supposed to be amazing*). But, this can be corrected using Eqn 9.7 after h has observed some data over the correct correlation between $x, c_s(x), y, f(x), f(c_s(x))$ and y', especially if there exists a lower-complexity function mapping between them - for instance, adding the phrase "supposed to be" may alter the label of a review.

This is similar to boosting [119] related methods where the original classifier f's errors on the counterfactuals is being learnt by the auxiliary classifier h. This helps us understand why the pairwise counterfactual classification task might be easier and perform better than simply annotating the counterfactual example $c_s(x)$ using the original classifier f. We can draw parallels to boosting [119] and draw insights as to why the number of samples required might be less. We now proceed to how our methodology compares to baselines (including using f for annotation) on held-out counterfactual robustness and the impact it has on the original accuracy.

9.4 Evaluation

We evaluate on two NLP tasks, sentiment classification and question paraphrase, using two datasets namely the Stanford Sentiment Treebank (SST-2) [388] and the Quora Question Pair (QQP) [191, 422]. Below, we briefly explain the problem set up in both datasets, how the counterfactuals are generated in each and the corresponding counterfactual datasets across which we evaluate counterfactual robustness.

9.4.1 Counterfactual Generator: Polyjuice

We use a general purpose counterfactual text generator called Polyjuice [442], which extends CheckList [357], that has shown promise by improving diversity, fluency and grammatical correctness as evaluated by user studies. It covers a wide variety of commonly used counterfactual types including patterns of negation [214], adding or changing quantifiers [128], shuffle key phrases [461], word or phrase swaps which do not alter POS tags [367] or parse trees [438], along with insertions or deletion of constraints that do not alter the parse tree [282]. Specifically, we use 8 prompts or types of counterfactuals - negation, quantifier, lexical, resemantic, insert, delete, restructure, shuffle; in Polyjuice to generate the augmented dataset. Other text generative models like [468, 214, 202] that improve adversarial robustness or like [220, 79] that allow controlled generation could be used as well.

9.4.2 Tasks and Datasets

Stanford Sentiment Treebank: We use the sentiment analysis dataset SST-2 [388] which assigns a binary sentiment (negative/positive) to a sentence mined from RottenTomatoes movie reviews. The corresponding counterfactuals are generated using the Polyjuice generator [442]. The original dataset contained 4,000 samples, while the counterfactual dataset had 2,000 samples

with human labels against which we evaluate. We show a sample of the dataset in the following:

Positive: A dog is embraced by the dog
Negative: A dog is not embraced by the dog

Quora Question Pair: In the QQP dataset [191, 422], given a pair of questions, the task is to predict if they are semantically equivalent, hence marked as duplicate. Here, again the second question is modified by Polyjuice [442] as per the templates used for the SST-2 dataset including negation, insertion, deletion, rephrasing, etc, out of which 1,911 samples were human annotated for evaluation. The original dataset had 20,000 samples.

Duplicate: How can I help a friend experiencing serious depre
who is in depression?
Non-duplicate: How can I help a friend experiencing serious de
a friend who is in depression?

9.4.3 Baselines

We now briefly describe five different baselines used to generate the labels of counterfactual augmented data (Y'_g), given access to a small number of annotated labels Y'_a.

- **No-cda**: f without any counterfactual data used for robustness.
- **Label-invariant (invariant)** : the labels of the counterfactual examples are assumed to be the same as the corresponding original sentence: $y' = y$.
- **Trust**: we trust the classifier f to annotate the counterfactual labels $y' = f(c_s(x))$ - a form of semi-supervision based on the existing base classifier.
- **Weighted-trust** (**w-trust**): the label of the counterfactual example is computed via the maximum score weighted by the confidence score of the classifier f on the pair for a label $l : p_l(x)$ such that $y' = \arg\max_l p_l(x) \cdot p_l(c_s(x))$.
- **Random**: In order to understand the importance of the counterfactual sentences used in the pairwise classifier, we also evaluate against a classifier which takes two randomly paired

sentences from the original dataset as input and predicts the second label given the label of one sentence.

- **Training**: we only use those counterfactual examples with human-annotated labels (X'_a, Y'_a) and drop all other counterfactual examples.

For all these baselines as well as our proposed methods, we use the RoBERTa [255] fine-tuned model as the choice of classifier f, and a corresponding pairwise fine-tuning task using RoBERTa [1] for the auxiliary pairwise counterfactual classifier h.

9.4.4 Experiment Setup

In both datasets, we have a small number of counterfactual human annotations available (SST-2: 2,000; QQP: 1,911) [442]. We divide these examples into two sets, one for training and annotating using h, and another held-out test dataset used to compute counterfactual robustness of f. The former dataset is used for fine-tuning f for counterfactual robustness, while the latter is used only as a held-out test set. In the SST-2 dataset, this means we split out 1,000 samples for training/annotation and 1,000 as the test set, while in the QQP dataset, we use 1,000 samples for training/annotation and the remaining 911 samples for testing counterfactual robustness. However, our aim is to use a minimal subset of the 1,000 samples available for training the base classifier directly. Instead, we use a smaller training dataset (say 100) to train our pairwise classifier which in-turn can then *artificially* annotate the remaining (say 900) samples. The combination of these (sum to 1000) will then be used to train the base classifier. Thus, in all our experiments, the number of counterfactual samples available to the base classifier to train on remains the same, although at different levels of human labeling costs.

The classifier f is first trained on the original classifier and then fine-tuned on the counterfactual dataset. We also perform 10 random initializations of the model f and h and a 10-fold

[1] huggingface.co/roberta-large-mnli, textattack/roberta-base-SST-2, ji-xin/roberta_base-QQP-two_stage

cross-validation split on the training/annotation data, thus report the mean and standard error bounds $\sigma/\sqrt{n}$ over $n = 1000$ runs for each model-based annotation and training for counterfactual robustness. We used the standard hyperparameters provided 1 for training f on (X, Y) and the hyperparameters for fine-tuning f on (X'_t, Y'_t) include learning rate of $5e-5$, batch size of 16 and a sequence length of 120 for 20 epochs. The pairwise counterfactual classifier's hyperparameters were chosen after a grid search to have a learning rate of $5e-4$, batch size of 32 for 50 epochs, sequence length of 240 including the original label and classifier predictions with special marker characters. While the base classifier f is trained on contextual embeddings of the sentence(s), h is trained by further augmenting the original and counterfactual sentence embeddings as input to RoBERTa followed by the base classifier's predictions separated by special delimiters [DEL].

To test the methodology on out-of-domain datasets, we test on sentiment analysis tasks in 3 class-balanced reviews datasets - IMDB movie reviews, Amazon reviews, and Yelp reviews [215]. The IMDB reviews (1,700) were collected by [214] through careful human elicitation to produce label varying counterfactuals of existing IMDB reviews. In the Yelp reviews [19], the task is to predict the ratings of 115,907 reviews on a scale of 1-5, and in the Amazon reviews [307], we evaluate on the 57,947 reviews in the clothing product category. Each one of these datasets were not used for training either the base classifier or the pairwise classifier, and the training relies solely on the SST-2 dataset. So, we can measure the generalizability of the pairwise classifier based data augmentation methodology.

9.5 Results

9.5.1 Improving Counterfactual Robustness

To demonstrate the effectiveness of our proposed methods: pairwise-counterfactual (**PC**) and classifier-aware pairwise-counterfactual (**CAPC**), we perform counterfactual data augmentation

using 10% counterfactual examples with human-annotated labels as well as 90% counterfactual examples (a total of 1,000 samples), whose labels are predicted using each method. The error rate on the hold-out counterfactual examples (referred as robustness) as well as on the original test set are shown in Figure 9.2.

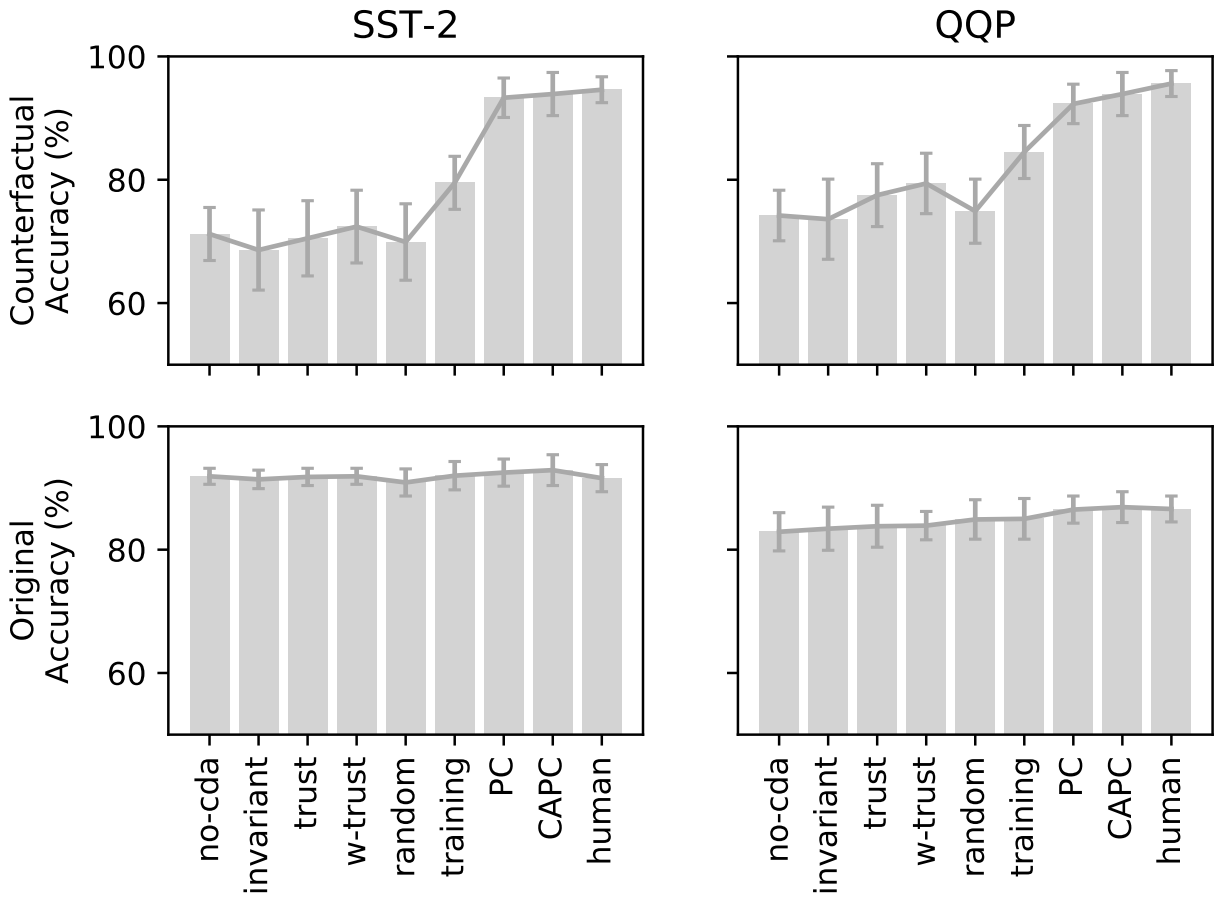


Figure 9.2: (a) Robustness: (first row) Training on 10% of human-annotated counterfactuals, and annotating the rest using the auxiliary classifier, we achieve a comparable improvement in robustness (lower error rate) for both Stanford Sentiment and Quora Question Pair datasets; **(b) Accuracy:** This improvement in robustness does not sacrifice the accuracy on the original held-out dataset.

We can clearly see that (1) the error rate of our proposed methods: **PC** and **CAPC** both significantly outperform other five baselines on models' robustness. (2) Comparing **PC** and **CAPC**, we can see that **CAPC** performs slightly better than **PC**. This indicates that the prediction of the original classifier $f(x)$, $f(c_s(x))$ does provide additional information to help with labels prediction. (3)

In addition, we also compare our methods with the extreme case that all the counterfactual examples (100%) are provided human-annotated labels, denoted as (**human-labels**). Surprisingly, our methods, which only use 10% human-annotated labels and predict the labels for the other 90% counterfactual data, achieve comparable performance in improving models' robustness. This sufficiently supports that our proposed methods can effectively predict the labels for counterfactual examples. (4) Looking at the error rate on the hold-out original test set, all the methods share a similar performance on SST-2 and our methods are better than other baselines and comparable to human-labels on QQP.

9.5.2 How much human-annotated data do we need?

To understand the impact of the training data provided to the auxiliary classifier h, we increased the % of data Y'_a provided to the classifier. While this increases costs of annotation, it is important to understand the headroom improvement in counterfactual robustness one would get had they opted for complete human-annotation. Figure 9.3 shows that across both datasets, the improvement in accuracy and robustness in providing more human annotations to train h : CAPC and subsequently training the model f : RoBERTa-{SST-2, QQP} is not significant and hence further demonstrates that, with just 10% of the augmentation dataset, we can already achieve an improvement comparable to a fully human annotated dataset. This further confirms our method can achieve high *sample efficiency* in improving models' robustness.

9.5.3 Generalization across Counterfactual Types

We evaluate the generalization of our pairwise counterfactual classifier h by ablating one counterfactual type (e.g negation, quantifier, etc) at a time during training h, but still annotate them to generate the augmented training data for f. The results are shown in Table 9.1 (rows 2-7). We see that our approach outperforms existing baselines on counterfactual robustness. This further

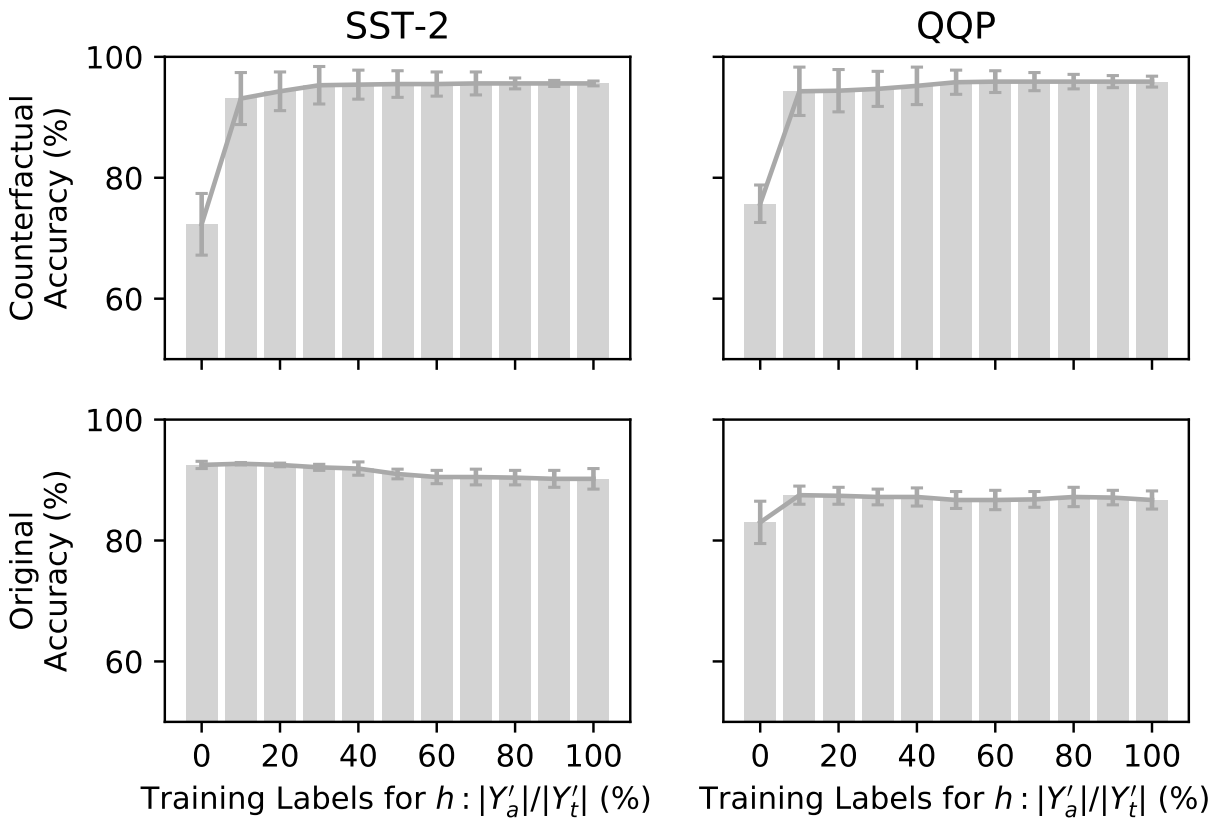


Figure 9.3: Impact of training size: As the number of samples $|Y'_a|$ increases more than 10%, there is not much headroom in counterfactual accuracy, and does not significantly impact the accuracy on the held-out original test dataset on both SST-2 and QQP datasets (overlapping error bounds).

Sliced Error when Counterfactual Type is Ablated %								
Model	**negation**	**quantifier**	**lexical**	**resemantic**	**insert**	**delete**	**restructure**	**shuffle**
CAPC-no-ablation	3.20	2.01	1.94	2.00	2.10	2.45	3.32	4.03
Generalization when counterfactual type is ablated from training h								
invariant	14.62	4.82	4.32	3.10	7.72	7.83	6.48	9.24
trust	12.96	4.15	4.73	3.00	4.95	12.49	3.74	9.02
w-trust	5.09	3.55	8.91	10.60	7.72	5.57	10.51	10.60
random	4.74	4.04	6.92	2.22	7.42	5.55	5.72	4.96
PC	4.50	5.35	**2.73**	3.20	**2.12**	**2.13**	5.30	5.10
CAPC	**4.04**	**2.20**	4.76	**2.10**	4.56	4.67	**3.56**	**4.50**
Generalization when counterfactual type is ablated from training h and f								
PC	7.02	7.40	4.63	5.35	2.42	2.54	6.85	9.34
CAPC	11.17	13.02	7.55	13.33	4.98	5.76	10.77	9.01

Table 9.1: Generalization of Counterfactual Types: Increase in error rates (%) of different counterfactual sentence types shows that our approaches CAPC and PC generalize better when those types are held out during training h. However, when we ablate the counterfactual type both while training f and h, our approaches perform comparably to the baselines. This shows that h does not just memorize the templates, but training on diverse counterfactual types is important for robustness

indicates the importance of learning a counterfactual classifier which captures patterns of label invariance that generalizes across counterfactual templates. Finally, we evaluate if our generated augmentation dataset can be used to improve *unseen* counterfactual types - ablated while training both h and f. While this is not the goal of our paper, it is useful to understand what types of counterfactuals are captured by our generator and if any overlap between the types of counterfactuals is leveraged. Table 9.1 (rows 8-9) shows that our approach is comparable with baselines (rows 2-5 in Table 9.1) when a specific counterfactual type is ablated completely from the data augmentation pipeline. This is consistent with existing work [198, 186] and further highlights the need to incorporate diverse types of counterfactuals to perform data augmentation.

9.5.4 Checklist Evaluation

To further validate that the generated labels by our auxiliary model can be used for other tasks, we evaluate it against the labels in CheckList [357] which capture other types of counterfactuals. We measure the *Absolute Failure Gap*: $|\epsilon - \epsilon_a|$ computed as the difference between the true error rate ϵ and the error rate as reported by using our augmented dataset ϵ_a while evaluating the models and

Test error rate %			
Model	**IMDB**	**Yelp**	**Amazon**
no-CDA	9.2	15.7	20.0
invariant	11.3	15.9	21.5
trust	9.3	15.8	20.5
w-trust	9.2	15.5	20.2
random	10.4	16.3	23.8
PC	8.0	14.3	18.1
CAPC	**7.2**	**13.1**	**17.2**
domain-trained	6.7	13.0	16.7

Table 9.2: Out-of-domain reviews: Using data augmentation with SST-2 counterfactuals from the Polyjuice generator and classified using CAPC performs comparable to a model trained on within-domain data.

tasks in the CheckList dataset. In Figure 9.4, we see that even when the training data provided to the auxiliary classifier is synthetically made explicitly label-invariant (90%), evaluating against counterfactuals with minimal label-invariance (10%), our model generalizes with a lower failure gap than other augmentation approaches. However, on the original Checklist dataset there is no significant improvement in failure gap compared to reporting the failure gap just on the training data alone.

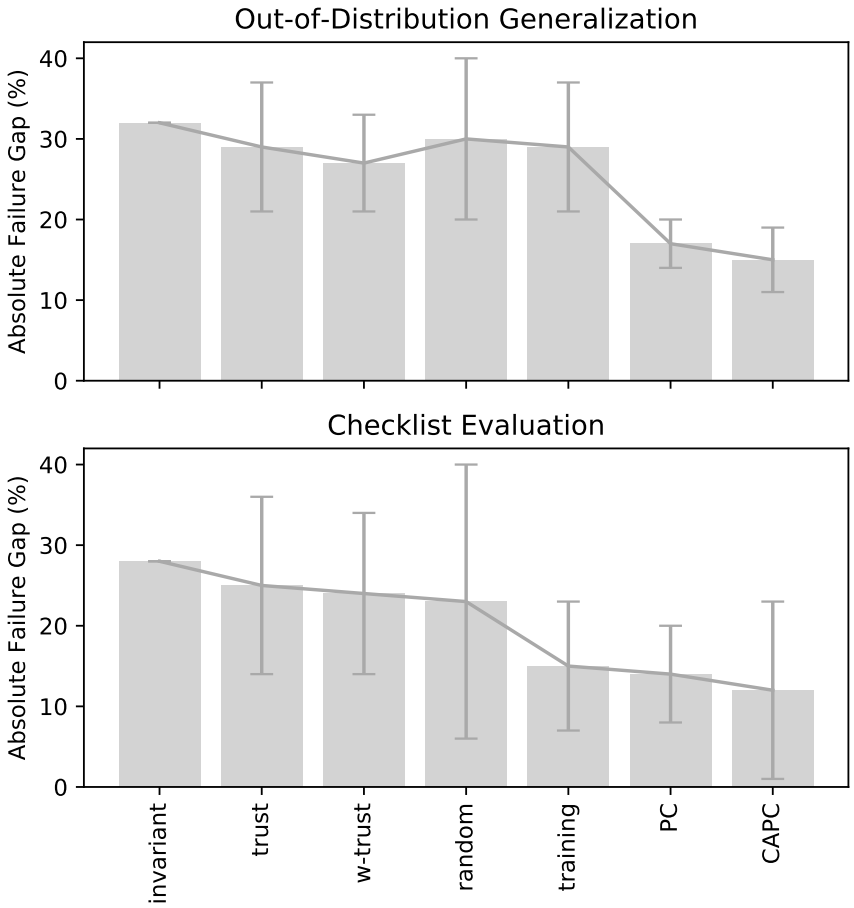


Figure 9.4: Checklist Evaluation - (a) Out of distribution data: Our methods perform well over different label-invariant distributions with 90% counterfactual label flips ($y \neq y'$) in the Checklist dataset even when the training distribution has only 10% counterfactual label flips; **(b) Model Comparison**: However, on the original Checklist dataset [357], we achieve a comparable failure gap with the golden error rate to other model-based annotations

9.5.5 Out-of-Domain Reviews

To validate that the counterfactuals we augment through our pairwise classifier's annotations has generalizability to out-of-domain datasets, we evaluate the reduction in error rates of the base RoBERTa model when they are trained on the pairwise classifier's data augmentation in Table 9.2. In the IMDB reviews dataset, we see an improvement in error rates from 9.2% without data augmentation to 7.2% through CAPC. This out-of-domain error rate is comparable to the error rate obtained by the model trained by [214] after incorporating samples from the counterfactuals drawn from the same distribution as part of the training (6.7%). In the Yelp reviews too, we see a reduction from 15.7% to 13.1% whereas other baseline approaches lead to an increase in error rates. Finally, in the Amazon reviews, the CAPC approach (17.2%) outperforms the baselines and is comparable to the augmentation from the training split from the Amazon reviews (16.7%). Each of these improvements have to be viewed with the context that it was achieved in a more sample efficient manner (1,000 counterfactuals generated from the original SST-2 dataset by Polyjuice) as compared to the in-distribution training approach, where the training data has 3,400 samples from their own respective datasets. This further confirms that training on augmented counterfactuals using a generator and pairwise classifier approach is comparable to human-annotated samples from other domains, while providing us the ability to scale both in terms of domain generalization as well as labeling efficiency.

9.5.6 Discussion

The need to ensure that natural language models predict reliably when sentences are perturbed in specific syntactic and semantically meaningful ways, beyond the observed training dataset is well established. Even though a checklist based framework introduces many constraints at once, it is important to ensure that enforcing one does not counter another counterfactual behavior. We now discuss how future work can build on top of our framework to overcome these limitations.

Importance of diverse templates While we show generalization across label variance in templates, we cannot guarantee that by learning solely on label invariant counterfactuals, our classifier can generalize over label modifying counterfactuals. Here, it is important to analyze counterfactual generators as to what type of sentences they generate and how it might be relevant to downstream tasks. While generators like Polyjuice [442] have been evaluated for fluency, diversity, etc., there is a need to evaluate them within the context of a task and its labels.

We improve what we measure One thing to note is that the set of counterfactuals we improved robustness over is limited. Our analysis indicates the need for more diverse counterfactual types that require a case-by-case contextual understanding. We show that adding more counterfactual types can be done in a sample efficient manner by using a generator trained to produce counterfactuals and a classifier which labels them by training on a small set of human annotations. Having more automated ways to improve robustness of natural language classifier would be an interesting future direction.

Using crowdsourcing efficiently The gains in robustness shown in Figure 3 and Table 2 further illustrate the need to dataset generation in an efficient manner. As future work, one can also look towards an efficient crowdsourcing strategy that minimizes the gain provided by the pairwise classifiers as each sample in the annotated dataset provide a unique and diverse counterfactual or a combination of counterfactual patterns that are not immediately evident from the previous set of samples. This can include prompts such as this sentence is similar to the ones already in the dataset and could encourage the human annotator to provide a sample different than already available.

Model Cards and Datasheets Each of the individual augmented counterfactuals generated from Polyjuice need to be incorporated in the Model Cards [295] under training sections, along with the intended use of such a dataset in the datasheet [134]. However, with such a generated dataset and an auxiliary classifier working in combination to produce the labels, the intended use of this combination is expected to be restricted for improving robustness.

9.6 Conclusion

Counterfactual Data Augmentation approaches have been extensively used to train for counterfactual robustness. As the types of counterfactuals - both label-invariant and label-modifying, over which to evaluate natural language models increase, there is a need to adopt a methodology that can scale with increasing types of counterfactuals. We overcome a significant challenge in doing so, by learning an auxiliary pairwise counterfactual classifier that leverages the patterns of counterfactuals produced by vairous generative models. Using only a small amount of human annotated counterfactual samples, we demonstrate that our method can produce a dataset that improves counterfactual robustness comparable to that of a fully human-annotated dataset.

Part IV

Domain Faithful Evaluation

10 Transparent demographic group trade-offs in Credit Risk and Income Classification

10.1 Introduction

In recent discussions of ethical ML algorithms, evaluating fairness has been frequently predicated on defining constraints based on specific protected attributes, such as race or gender [412, 39]. These attributes should **not** demonstrate conditionally discriminative behavior while learning classification targets. If care is not taken in the construction of an ML model, works such as [465] and [105] have shown that inequalities in underlying data distributions can be amplified in the predicted output, leading to runaway feedback loops. Recent works [217] have argued that examining the intersectionality of multiple protected attributes is crucial for establishing coherent standards of fairness. However, real-world data sub-populations often display varying underlying sampling distributions, bias and noise. We argue that principles towards *fair* ML should encourage transparency in the trade-offs between demographic group accuracy in a classification task and at a minimum be able to reflect their true underlying population distributions. At a fixed sample size, as the number of protected attributes increases, the intersectional subgroup populations tend to decrease in size. In these scenarios, it is evident that any classifier which

does not perform worse on all groups can never be *fair* [288]. Hence, in this paper we aim to understand the research question of how the trade-offs between demographic groups affect the evaluation of different methodologies proposed that are aimed to improve classification accuracy. To quantify this, we look to the rich literature of "individual fairness" which defines fairness with respect to a similarity metric between two individuals and enforces that similar individuals are treated similarly, within an error bound [100, 99, 217]. We find this definition to be useful in allowing us to continue to ensure that minority demographic group populations perform at their best accuracy while ensuring that majority demographic groups do not suffer a large decrease in group-level accuracy.

Using this transparent Pareto-principle of Efficiency [144], popular in social welfare and economics, we argue that trade-offs between demographic group accuracy undertaken by ML algorithms in high-stakes applications like credit risk and income classification [88] should be made transparent in order to be examined against socio-technical norms in that application domain [373, 158, 266]. We have been motivated by the insight that many fairness problems in existing classification tasks for specific subpopulations can be remedied by controlled data collection, subject to ethical considerations [52, 59]. As such, we suggest that in the spirit of achieving fair outcomes, when learning on datasets with varying demographic group sample sizes, how we weigh the loss suffered by each demographic group can be a critical choice and should be transparent.

In the domain of credit risk assessment, the trade-off between the accuracy of demographic groups has implications on financial justice across demographic groups. For example, older married male individuals have better accuracy than younger single female individuals for credit risk assessment. This means that even a seemingly group-blind ML algorithm can have significantly different accuracy across demographic groups. Similarly, in the income classification task, Caucasian male individuals have much better baseline accuracy than non-Caucasian female individuals in the United States. Therefore, to build transparent and fair ML systems, we show that

the trade-offs between these demographic groups cannot be avoided, but rather should be an integral part of the transparent design of any socio-technical ML system. We illustrate one such transparent trade-off mechanism by arguing for efficiency based on the Pareto principle, where degradation in the accuracy of one group should not occur without improving another group's accuracy. In this paper, we compare our transparent Pareto-principle based trade-off with several other strict equality-based constraints and demonstrate an increase in 9.5% and 9.6% overall and group-level accuracy respectively on both the credit risk and income classification tasks.

10.2 Motivation: Trade-offs in the real world

10.2.0.1 COMPAS

A ML model (COMPAS tool) was used for determining the risk of recidivism in Broward County, Florida, USA. ProPublica [**propublica**] found in an independent investigation involving 18,610 people over 2 years that black males were twice as likely to be misclassified by the model as high risk as compared to white males. This scenario highlights the critical need for auditing existing decision-making systems (including the ones based on human experts) and understanding the trade-offs made in their design. In such a high stakes scenario, ideally, a decision-making system that achieves the highest group level metrics (such as accuracy) is required. By incorporating inductive biases based on racial and social justice, one could hope to achieve the end objective of improving the Pareto front transparently. If we do not attempt to evaluate and discover Pareto efficient classifiers, a domain expert choosing a classifier might end up making trade-offs of accuracy and fairness among inefficient classifiers.

10.2.0.2 Gender Shades

Certain image recognition models were discovered to have lower accuracy for one particular group (darker females) than other groups in the Gender Shades project [116]. The intervention

undertaken to resolve this discrepancy involved collecting better data for the poor performing group (females with darker skin tone). The progress from such interventions amounts to discovering better group accuracies on the Pareto frontier, as opposed to restricting the models to strict equality among groups. Here too, the authors of the project, Buolamwini and Gebru, advocate for a complete ban of ML models for facial recognition tasks since these models are not advanced enough to perform with high accuracy on all groups independent of skin tone and gender, without encoding spurious correlations. Hence, a ML model needs to be transparent in the trade-offs that it implicitly makes to gain socio-technical acceptance in the real world.

10.3 Transparent Trade-offs

The Pareto frontier has been used to characterize the trade-offs between more than one dimension in multiple objective learning [7, 353]. It characterizes solutions such that no point on the Pareto curve dominates another point on all the dimensions across which we measure an objective. Evaluating the Pareto curve for any ML classifier can be critical in making transparent trade-offs between demographic groups [3].

10.3.1 Pareto front in ML based models

In our analysis of the German Credit and Adult Census Datasets, we take an example of a feedforward neural network model with up to 3 layers with each layer containing 256, 128 and 64 hidden units respectively. We then perform a sweep of the hyperparameters by varying the depth of the neural network, learning rates and L1 and L2 regularization parameters [304], and the training data made available to train the network (specific demographic groups versus the entire dataset). Each network was trained multiple times with randomized seeds for initializing the parameters of the network. This gives us a wide range of group-level accuracies along each of the demographic groups we slice the accuracy of the model. We then constructed the Pareto front of these

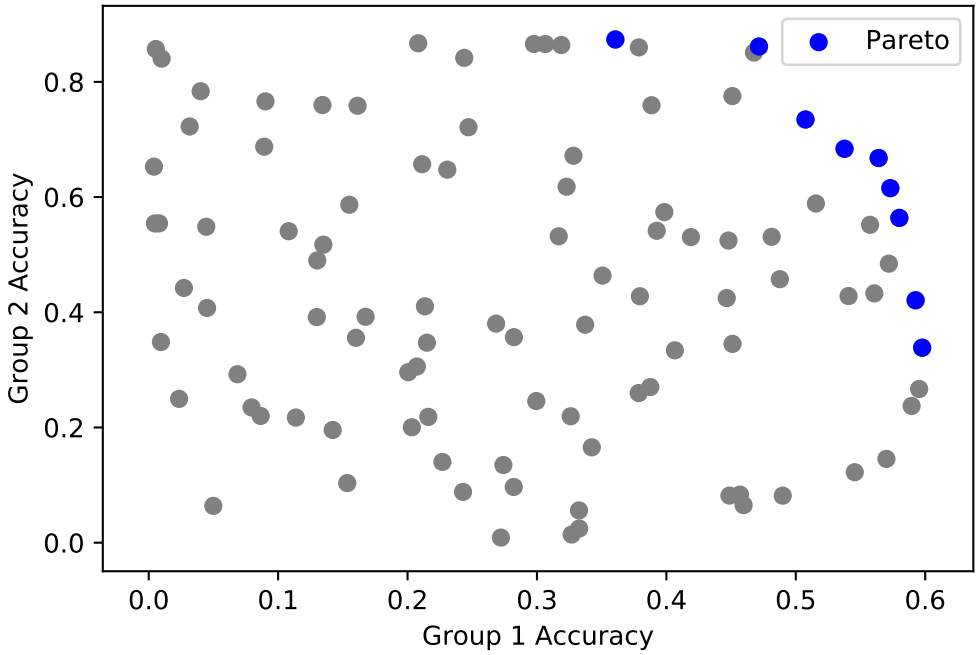


Figure 10.1: An illustration of a two group-setting plotting group-level accuracy and its corresponding Pareto front (in blue) shows that demographic group trade-offs are implicit and unavoidable in ML systems

group-level accuracies after varying the hyper-parameters, with each group corresponding to a dimension of the Pareto front. Note that visualization of the Pareto front can be tricky, given that in most real-world applications, the demographic groups are more than three. Hence, we need a principled approach using which a domain practitioner can argue about their choice of a specific classifier on the Pareto front. In figure 10.1, we see that in simulated data with two demographic groups, a domain expert can trade-off one group's performance with another by choosing different points on the Pareto front. Also, we can see that a trade-off is inevitable unless we assume that the Pareto front exactly intersects with the hyperplane where all demographic groups perform equally (x=y in case of two dimensions).

10.3.2 Pareto Trade-offs

Having established that a trade-off between group-level accuracy should be conducted on the Pareto front, we now provide an example where such a trade-off is transparent and based on a Pareto Efficient and fair principle. In this principle, a domain expert might choose a classifier where each group's performance sacrifices accuracy equally. For example, in the German Credit risk assessment task, older male individuals can achieve their best accuracy of 91% among all the points on the Pareto front, whereas younger female individuals can achieve only 73%. In this case, the Pareto-based trade-off would advocate for a classifier that achieves 89% and 71.4% on the two groups respectively, each of them about 2.2% below (Pareto Loss) their respective optimal choices on the Pareto front. This choice is different than the one a domain expert would choose based on the principle of strict equality or Demographic Parity [168] between the groups (both groups at 73%, i.e. zero Parity Loss). We acknowledge that both of these choices might be valid in different contexts based on the principles the corresponding algorithmic decision-making system prescribes. But, the choice needs to be transparent and cannot be masked behind the objective of minimizing overall classification error. This transparency allows people who apply for credit to contend the trade-offs and the corresponding principles in automated decision-making systems. Hence, with transparency, the people who were previously left out of the decision-making systems' design can be involved and provide them the ability to appeal the trade-offs made by such ML models.

10.4 Evaluation

10.4.1 Baselines

We compare our transparent trade-off approach with optimization techniques that use fairness constraints such as Equality Constraint [465], Adversarial [37], and Min-Max fairness [275]. Zhao

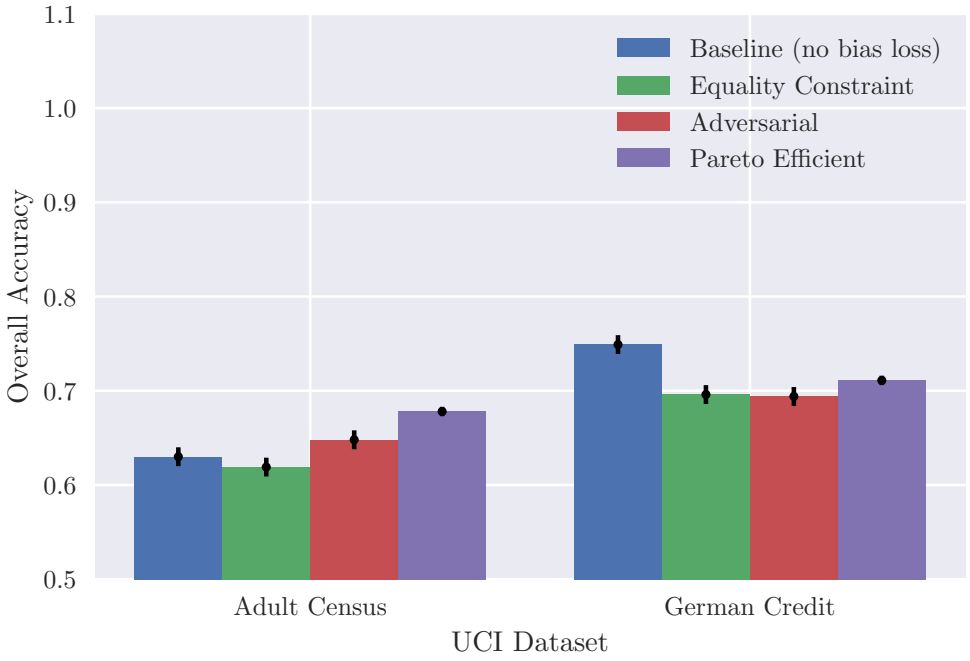


Figure 10.2: Comparison for 2 UCI datasets showing that the pareto-based transparent trade-off achieves better overall accuracy than other fairness constrained classifiers.

et al. [465] aim to lower the sum of absolute discrepancy of all group accuracy from the overall accuracy (Parity loss), while Beutel et al. [37] adversarially attempt to nudge the classifier such that it cannot predict the protected attributes. Martinez et al. [275] aim to maximize the accuracy of the least performing demographic group.

10.4.2 UCI Adult Dataset

The UCI Census Adult dataset focuses on the prediction of income as a binary variable (> \$50K, <= \$50K) based on demographic information. Protected attributes selected are gender and race and are denoted as binary categorical variables. We consider the 4 groups at the intersection of the protected attributes, to overcome the limitations of group fairness as outlined in [218]. The dataset has 48,842 instances out of which 20% is held out as test data, while the remaining is

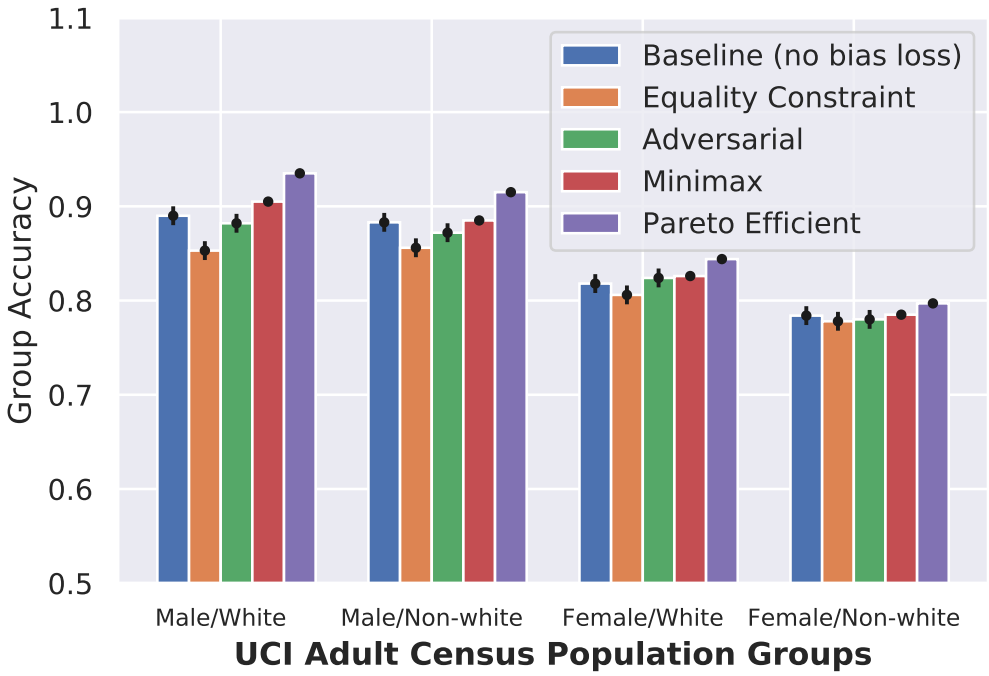


Figure 10.3: Group accuracy comparison shows that we achieve Pareto dominating group level accuracy for all groups in UCI Adult dataset.

used for training and cross-validation. There are 14 attributes out of which 6 are continuous and 8 variables are categorical. Table 11.2 shows the Pareto loss, i.e how much each group deviates from the pseudo-optimal of the respective group for the UCI Census Adult dataset. Based on the Pareto principle, we were able to choose an optimal point on the Pareto front that ensured that each of the demographic groups perform optimally. In our transparent trade-off on the Pareto front, each of the groups has better individual accuracy than the other approaches and thus better overall accuracy as shown in Fig 10.2. Fig 10.3 demonstrates that our approach arrives at a better classifier on all demographic groups. Some groups even exceed the baseline accuracy (computed using the average of all unconstrained optimization results) due to an extensive swap of the hyperparameters and transparently choosing the Pareto optimal classifier.

Model	FPR	FNR	Parity Loss	Pareto Loss
Baseline (no bias loss)	0.253	0.747	0.199	0.016
Equality Constraint[465]	0.283	**0.712**	**0.167**	0.133
Adversarial [37]	0.224	0.769	0.226	0.077
Min-max [275]	0.202	0.773	0.218	0.075
Pareto Efficient	**0.165**	0.830	0.250	**0.000**

Table 10.1: Comparison of test losses in UCI Adult dataset - False Positive Rate (FPR), False Negative Rate (FNR), Parity and Pareto Losses. Our Pareto-based trade-off has no difference as compared to the Pareto optimal group-accuracy, while [465] minimizes Parity loss.

10.4.3 UCI German Credit Dataset

The UCI German Credit risk assessment dataset involves predicting credit type as a binary label (good or bad) from demographic information where the protected attributes selected are age, gender and personal status. Each of these protected attributes is binarized and the intersection of these 3 attributes is considered as the groups in our study. There are 1000 instances in the dataset with a total of 20 categorical attributes. We hold out a random 20 % as test data over which we present the results. The evaluation of this dataset is determined by a cost matrix where the false positives are considered 5 times more costly than a false negative. The final accuracy reported takes this into account. Similar to the UCI Adult Dataset, in Figure 10.4, we see that choosing a point based on our Pareto principle, we increase the group-level accuracies as compared to the equality constraints [168], adversarial loss [37] and minimax [275] optimization techniques. The 5 groups (out of the total 8) are shown in the UCI German Credit Dataset, as the rest of the groups do not have enough samples (< 100).

10.4.4 Sample Size Inconsistencies

The use of explicit demographic attributes in real-world scenarios is sometimes a hard constraint. One example is legislation enforcing fairness around disparate impact [409, 111]. In simplified examples, exploring the intersectionality of protected attributes may be appropriate. For example,

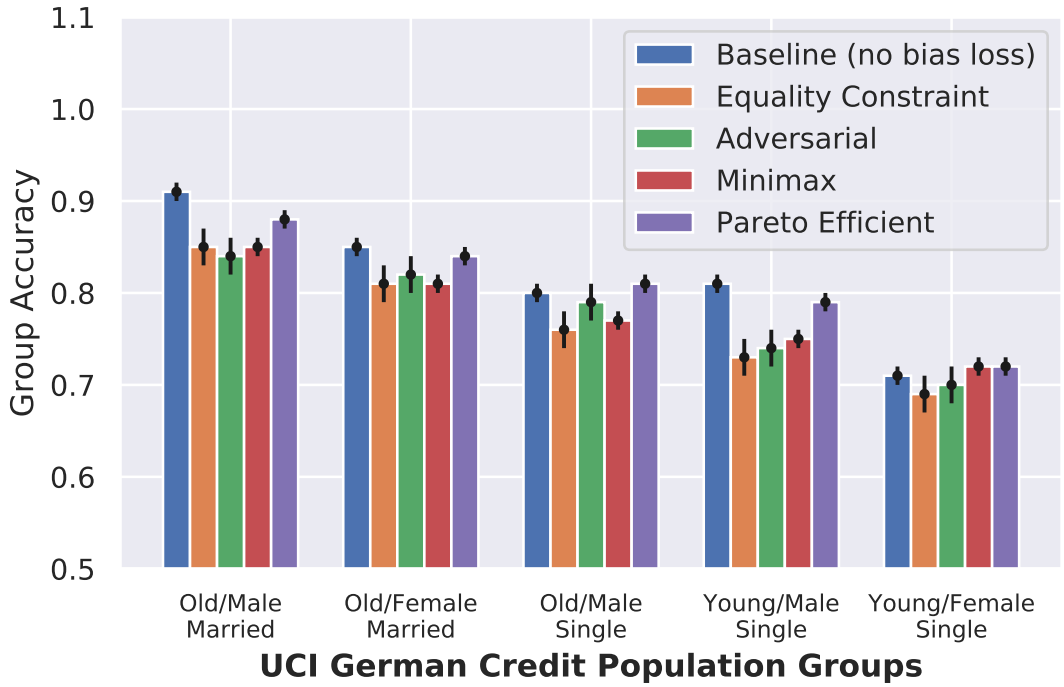


Figure 10.4: Group accuracy comparison showing we achieve optimal group level accuracy for all groups in UCI German Credit dataset among constrained classifiers.

in this paper, we explore two gender and two race subgroups in the evaluation of the UCI Adult dataset, which translates to four separate groups. It is conceivable that in a real-world application, the intersection of gender and race subgroups could extend into **many** different groups. As the intersectionality of groups grows, a group's sample size will likely be insufficient. In the case of the UCI German credit risk assessment dataset, the attribute - marriage status, with five possible values, is treated as a protected attribute along with gender and age. However, in the dataset, there were **no** samples containing both the attributes of young, female and being married.

Despite the impossibility results of achieving fairness in the extreme case of subgroup sized one, there is still a need to highlight cases where simple (linear) models are inadequately applied in datasets with complex underlying subgroup distributions [**propublica**, 64]. The ability to transparently argue about the trade-offs made in designing the required model along with the

Table 10.2: Comparison of sample complexity ranking for Probably Approximately Metric Fairness with actual subgroup sizes of subgroups

#	Group	Complexity Rank	Sample Size (Rank)
UCI Adult Dataset			
1	Male/White	1	2,129 (1)
2	Male/Non-white	3	8,642 (3)
3	Female/White	4	2,616 (2)
4	Female/Non-white	2	19,174 (4)
UCI German Credit Dataset			
1	Old/Male	3	50 (1)
2	Old/Female	4	310 (3)
3	Young/Male	2	548 (4)
4	Young/Female	1	92 (2)

limitations of small sample sizes for certain demographic groups will guide the choices made by practitioners and ML researchers. Through our work, we see that even an ML model that does not explicitly perform a trade-off between demographic groups has already decided the trade-off implicitly.

Using the theory of sample complexity based on Rademacher complexity [297, 380], if we assume all the hypotheses are linear with VC Dimension d, we can rank the hardness of learning the target for each demographic group, and order them (Table 10.2 - higher numbered rank has higher complexity values). The sample complexity to learn a PAC algorithm which achieves error of less than ϵ with probability δ in all k subgroups is lower bounded by m:

$$m = O(\frac{ln^2(k)}{\epsilon}((d+k)ln(\frac{1}{\epsilon}) + kln(\frac{1}{\delta})) \tag{10.1}$$

In the UCI Adult Census dataset, the ordering of the actual subgroup sample sizes ($4 > 2 > 3 > 1$) reveals that new samples are needed to match the desired sample complexity ordering ($3 > 2 > 4 > 1$). Specifically, more samples for subgroup 3 (Female/White) need to be gathered than for subgroup 2 (Male/Non-white) to ensure the ordering of actual sample sizes aligns with that of the sample complexities. Similarly, in the German Credit Dataset, Table 10.2 shows disparity in

the order of the actual sample sizes ($3 > 2 > 4 > 1$) as compared to desired sample complexity ($2 > 1 > 3 > 4$). This implies that in the UCI German Credit dataset, more new samples from group 2 (Old/Female) than from group 3 (Young/Male) should be drawn for us to make a balanced and transparent choice while performing trade-offs. Similarly, more samples from subgroup 1 (Old/Male) need to be collected than from subgroup 4 (Young/Female) to remove any inversion in the ranking of complexities and actual group sample sizes to ensure that the trade-offs are not performed inefficiently due to insufficient sample sizes.

10.5 Conclusion

We advocate for transparency in the demographic group accuracy trade-offs in high-stakes real-world applications like credit risk and income classification tasks. We demonstrate that transparency in how we balance group-level accuracies can lead to better classifiers being explored on the Pareto front while improving overall accuracy too by 9.5%. Further, we caveat that trade-offs on demographic groups with smaller sample sizes should be taken into account and appropriate data collection exercises should be conducted. We argue that for the development of an ethical AI framework for policy and decision-makers, transparency in the group-level accuracy trade-offs is critical. Future work to extend this analysis to more complex ML models may provide principled standards for transparent trade-offs between groups in other application domains along with mechanisms to contest them.

11 Predicting Angiographic Disease Status: Drawing the line between demographically decoupled and jointly trained models

11.1 Introduction

Societal inequities have the real risk of being vastly exacerbated if machine learning algorithms do not take explicitly address issues of demographic inequity [412, 39, 465, 176, 46]. In the context of diagnosis of angiographic disease status, age and gender based demographic groups have been known to have prognostic differences in CT coronary angiography, with females below 60 years of age have the least predictive value [452]. Prior work have also shown that women have higher mortality from myocardial infarction, mostly at younger ages [386, 267, 341]. Given that different demographic groups based on age and gender have different profiles of heart disease, the problem of improving the predictive accuracy of diagnostics across such demographic groups has not been explicitly tackled as the primary objective. Instead, there is an emphasis on overall accuracy of patients when using Machine Learning (ML) based predictive models. In this paper, we define the notion of "Demographic Pareto Efficiency" (interchangeably referred to as pareto efficiency)

as a guiding principle for domain experts to choose diagnostic models that improve predictive accuracy of angiographic disease status across demographic groups based on age and gender; and provide a methodology that discovers a larger set of ML models that consistently improve upon the predictive accuracy for all demographic groups. Specifically, our methodology makes the choice between learning separate decoupled models, one for each of the group, and a joint model trained on all groups based the main outcome measure of demographic group-level accuracies.

Improving equity in health is well studied and various philosophical notions of fairness exist (distributive, procedural, etc.) [373, 158, 363, 266] and the appropriateness of each definition depends on the ethical context in which they are applied. Theoretically, in an equitable world of perfect data, a classifier with perfect diagnostic accuracy across all subgroup populations may be created. Due to a variety of reasons including historical injustices [159], sampling bias [56], selection bias [389], label noise, among others, group populations are often not fully represented in commonly used real-world health datasets [8]. With such skewed data, [288] has shown that an unavoidable trade-off exists between group fairness and accuracy. With this trade-off, domain experts have to choose between coupled (jointly-trained on all groups) and decoupled (one model per group) models based on how well they balance the demographic group accuracies. While the benefits of decoupled models are known theoretically when we have large and diverse datasets [101], the impact of such models on group-level accuracy in diagnosing the angiographic disease status in patients remains unexplored. We investigate the role of decoupled training across demographic groups based on age and gender in the UCI Heart Disease dataset. Inspired by social science and welfare economics literature [325, 81, 276] (see S.I for detailed related work), we propose a novel methodology that combines decoupled group-wise models and use them to guide a jointly trained model to achieve demographic pareto efficiency [144]. .

Demographic Pareto Efficiency [144] is achieved when no single group performance can be improved without the degradation in performance of another group. The set of all such group level performances when plotted in a multi-dimensional graph (one group's performance per dimen-

Optimization Objective	Operating Point
Overall Accuracy	$opt_b = (0.63, 0.77)$
Strict Accuracy Equality[465]	(0.60, 0.60)
Adversarial [37]	(0.73, 0.56)
Mini-max [275]	(0.68, 0.63)
Pareto Efficiency (Ours)	$PE = (0.71, 0.63)$

Table 11.1: Preferred classifiers and their demographic group-level accuracy based on different objectives in Fig 11.1.

sion), forms the Pareto frontier (like blue dots illustrated in a simulation shown in Figure 11.1). Ensuring that classifiers achieve Demographic Pareto Efficiency while balancing fairness constraints and accuracy has critical implications to the discussion about the unavoidable accuracy-fairness trade-offs in the real world [288]. For example, if domain practitioners are required to make a choice between two classifiers based on the demographic accuracy-fairness trade-off in predicting the Angiographic disease status, the comparison would be meaningful only if both those classifiers were on the Pareto frontier. Otherwise, the discussion of demographic accuracy and fairness trade-offs would be premature as there exists a third classifier which can achieve better group level accuracy and better medical outcomes. (e.g.: "Pareto Efficient Fairness" should be preferred over "Strict Accuracy Equality" in Table 11.1). Through our approach, we discover such Pareto efficient predictive models to be considered as candidates in determining the angiographic disease status of patients, and avoid unnecessary concessions in group level accuracy without significant degradation in fairness (demographic parity).

11.2 Background

Problem Definition: The angiographic disease status is defined as a binary label (diseased or not) based on the fact if there is more than 50% diameter narrowing in any of the major blood vessels in a patient (lmt, ladprox, laddist, diag, cxmain, ramus, om1, om2, rcaprox, rcadist). To

predict this angiographic disease status, we use 13 input attributes of the patient such as age, gender, chest pain type (typical angina, atypical angina, non-anginal and asymptomatic), resting blood pressure (mm Hg on admission to hospital), serum cholestrol (mg/dl), fasting blood sugar (binary >120 mg/dl), resting electrocardiographic results (normal, having ST-T wave abonormality, probable or definite left ventricular hypertrophy), maximum heart rate achieved, exercise induced angina (yes/no), ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and detection of thalassemia (none, major or reversible defect). For this heart disease dataset, different machine learning algorithms may be trained on the same training dataset of patients, and may obtain different test accuracies on demographic groups as illustrated in Figure 11.1. For example, when we plot each ML algorithm as a separate point, where the value along the x and y axis indicate the test accuracy over demographic groups A and B respectively, we see that some algorithms perform poorly on both groups (annotated as alg-1..5 in grey) as compared to other algorithms (annotated as opt_b, overall, PE, opt_a in blue). The blue line indicates the Pareto frontier of the demographic group test accuracies, which defines the set of optimal choices a domain expert would have if they were to optimize for demographic group test accuracies. Note that each algorithm on the Pareto frontier when compared with another algorithm on the Pareto frontier, performs better on one demographic group and poorly on the other, but never poorly on both the demographic groups. Hence any choice the domain expert would make among these ML algorithms would need to trade-off one demographic group's accuracy for the other. While this choice depends on the domain expert and the context in which they operate (other remedial or diagnostic measures incorporated for specific demographic groups), the key takeaway from this illustration is that optimizing for overall accuracy without consideration of the demographic groups may lead us to one of the points on the Pareto frontier, but may not always be the one that the domain expert would choose given all options on the Pareto frontier. *Hence, our primary goal in this paper is to discover all the machine algorithms on the Pareto frontier, so that the domain expert is given a*

rich set of algorithms with Pareto optimal demographic group accuracies to choose from. However, discovering the Pareto frontier is non-trivial as their discovery is driven by optimizing specific demographic group accuracies, while keeping the accuracy of other algorithms constant. Further, this problem is exacerbated by disparate sampling bias, epistemic, and aleatoric uncertainty about how angiographic disease status presents itself in different demographic groups. Thus, it is not known that a simplistic training objective such as improving overall accuracy is sufficient for discovering the full Pareto frontier. Thus, in order to discover the full Pareto frontier in a systematic approach, we present an iterative training methodology.

11.3 Methodology

We now formally define Demographic Pareto Efficiency and explain how to train a joint model that leverages the benefits of decoupled classifiers in discovering Pareto efficient classifiers on the Pareto frontier.

Definition 11.1. Demographic Pareto Efficiency: We introduce Demographic Pareto Efficiency as a set of classifiers with respect to groups (defined by sensitive attributes). Demographic Pareto Efficient classifiers are defined as the set of classifiers where there does not exist another classifier which has better performance (for a defined performance metric such as accuracy, TPR, etc) across all groups.

For groups $g \in |G|$, we denote $(f_1, f_2 ... f_{|G|})$ as the tuple of group performance metrics achieved by any Demographic Pareto Efficient classifier. The $f_1, f_2, .. f_{|G|}$ are group performance metric values that are to be maximized (e.g. accuracy, TPR). Formally, Demographic Pareto Efficiency states that no other classifier exists with performance metrics $(q_1, q_2 ... q_{|G|})$ such that $f_1 \leq q_1$ and $f_2 \leq q_2$ and .. $f_{|G|} \leq q_{|G|}$.

Definition 11.2. Pareto Loss: Pareto Loss, ϵ_g, for a group g, is defined as the relative difference between the performance of a classifier for that group f_g and the optimal performance for the

group f_{opt-g} across all discovered classifiers.

$$\epsilon_g = 1 - \frac{f_g}{f_{opt-g}} \tag{11.1}$$

While the above formulation of optimizing the Pareto loss can lead to multiple Pareto efficient decoupled and jointly trained diagnostic models of angiographic disease, the domain expert has to choose a single classifier among them post-training. Ideally, choosing the most desirable classifier is left to the end, once the complete Pareto front has been discovered. As discovering the Pareto frontier itself is our problem statement, this can lead to a deadlock condition, where an effective choice between decoupled and jointly trained models cannot be made without making choices that lead to better exploration at training time.

Definition 11.3. Pareto Efficient Fairness: We define a classifier as Pareto Efficient Fair (PEF) if it is Pareto Efficient and minimizes a weighted average of variance and absolute sum of the Pareto loss across groups.

The definition of Pareto loss requires us to know the true optimal performance per group f_{opt-g} a priori, which may not be possible. Hence, we use a decoupled classifier to estimate these optimal values at each iteration of training.

Pareto Efficient Algorithm: A heuristic pseudo-optimal group accuracy f_{opt-g} for each group g is formulated by training a decoupled classifier M_g to minimize the cross-entropy loss $\mathcal{L}_{ce}$ on samples in group g from dataset D [324]. We then iteratively update f_{opt-g} if a better group accuracy is evaluated by a jointly trained model M on a held-out test set using the *eval* function. A summary of the Pareto Efficient bias mitigation algorithm is presented in Algorithm 6, and the corresponding components are explained in detail below. This is an in-processing algorithm (as opposed to post-processing [440]) which trains a joint model M on all subgroups to minimize the Pareto Efficient fairness loss $\mathcal{L}_p$ in every batch by stochastic gradient descent. We strictly ensure that the mini-batch is representative of the group distributions by sampling group-wise batch

samples proportionately. Our algorithm explicitly achieves *potentially optimal* performance for each of the groups by explicitly recognizing these differences [249] as opposed to ones which do so implicitly [221]. Now, we formally define our fairness based Pareto loss function $\mathcal{L}_p$ used in each iteration of our algorithm.

Algorithm 6 Iterative Pareto Efficient Bias Mitigation

G: set of sensitive groups, D: dataset, D_g: data of group $g \in G$
for $g \in G$ **do**
  $M_g = \arg\min \mathcal{L}_{ce}(D_g)$
  $f_{opt-g} = \text{eval}(M_g, D_g)$
  $f_g = \emptyset$
end for
while $\exists g \in G, f_g = \emptyset \vee f_g > f_{opt-g}$ **do**
  $f_{opt-g} = \max(f_g, f_{opt-g}), \forall g \in G$
  $train(M, \mathcal{L}_p(D))$
  $f_g = eval(M, D_g), \forall g \in G$
end while
return M

Consider a set of Demographic Pareto Efficient classifiers T_{PE}, with each classifier $t \in T_{PE}$ containing a tuple of Pareto losses $\mathcal{E}_{t,G} = (\epsilon_{t,1}, \epsilon_{t,2}, ...\epsilon_{t,|G|})$. The sample variance of Pareto loss across groups is denoted by $\sigma^2_{t,G}(\mathcal{E}_{t,G})$. The goal is to find the Pareto Efficient Fair classifier $t_{PE-fair}$ that minimizes the variance of Pareto losses among all groups. Since it is empirically difficult to find all the Demographic Pareto Efficient classifiers T_{PE} at each iteration of our algorithm, we relax this by approximating the Pareto classifiers as ones that have a low absolute sum of group Pareto losses ($\|\mathcal{E}_{t,G}\|_1$) among all classifiers $t \in T$. Since the classifier with the lowest absolute Pareto loss may not equate to the classifier that minimizes the variance of the Pareto loss across groups and vice-versa, we trade-off these two minimization criterion using a Lagrangian factor α in the *Group Pareto Loss* as follows:

$$t_{PE-fair} = \arg\min_{t \in T_{PE}} \sigma^2_{t,G}(\mathcal{E}_{t,G}) \tag{11.2}$$

$$\approx \arg\min_{t \in T} \alpha \|\mathcal{E}_{t,G}\|_1 + (1-\alpha)\sigma^2_{t,G}(\mathcal{E}_{t,G}) \tag{11.3}$$

When $\alpha = 0$, the variance of Pareto loss is minimized, whereas, when $\alpha = 1$, we minimize the absolute Pareto loss. In all our experiments, we chose $\alpha = 0.5$ after cross-validation, however the domain expert in the angiographic disease diagnoses might chose another value based on the trade-off between variance and absolute sum of Pareto losses. By making this choice explicit, we can demand transparency from practitioners deploying diagnostic ML models about the trade-offs they made. A high α would force that each demographic group be as close as possible to it's optimal performance, whereas a low α would enforce that each group suffer similar Pareto losses as compared to their optimal group performance.

Augmented Pareto Loss: We now generalize our definitions for any binary diagnostic model. Here, the minimization criterion of the Group Pareto Loss, but we minimize the group Pareto Loss over the parameters of the binary classification model using stochastic batch gradient descent. The Group Pareto Loss is augmented with an appropriate loss weight (λ) via the Lagrangian dual formulation similar to [103]. As an example, the standard cross-entropy classification loss: $\mathcal{L}_{ce}$ [271] can be augmented to yield the Pareto Efficient Fairness Loss: $\mathcal{L}_p$. The penalty term weighted by λ is used to ensure that maximum overall accuracy can be achieved while minimizing a combination of the absolute Pareto loss and its variance. After cross-validation, we set $\lambda = 0.1$, but here too the domain expert might choose based on external factors that impact the relative weight of overall as compared to group-level accuracy. (see S.I for detailed methods)

$$\mathcal{L}_p = \mathcal{L}_{ce} + \lambda(\alpha \|\mathcal{E}_G\|_1 + (1-\alpha)\sigma^2_G(\mathcal{E}_G)) \tag{11.4}$$

11.4 Results

Here, we predict health status as binary label (presence or absence of Heart Disease) using medical and demographic information, where we consider age (>60, <=60) and gender (male, female) to be the stratification variables. The intersection of these 2 variables are considered sensitive groups in our study.

The dataset consists of 920 patients from four hospitals of Cleveland Clinic Foundation; Hungarian Institute of Cardiology, Budapest, V.A. Medical Center, Long Beach, CA; and University Hospital, Zurich, Switzerland with a total of 75 attributes, out of which 13 attributes are used for predicting the binary label of angiographic disease status (0: <50% diameter narrowing, 1: >50% diameter narrowing). The number of samples in each of the four demographic groups Young/-Male, Young/Female, Old/Male, Old/Female are 550, 149, 176 and 45 respectively. We split the dataset into a 10-fold train/test random stratified splits (train on 9 splits, and test on the remaining split, repeated 10 times) based on the demographic groups to ensure that the training and test data are sampled from the same distribution and that all demographic groups are represented as per the dataset. We compare our approach with the scaled versions of group fairness [465] and [37] for groups. In [465], the authors optimize for overall accuracy in the constrained setting of ensuring equal false positive rates. The method is generally applicable to other measures of performance. For comparison, we implement an objective to maximize overall accuracy along with a Lagrangian relaxation which adds a penalty for parity loss (deviation from the overall accuracy) for each group.

This baseline scenario is equivalent to optimizing for balanced accuracy across sub-groups or assuming that perfect group-level performance can be achieved (accuracy of 100%). Instead, in our iterative approach, we use a per-group decoupled classifier's pareto optimal performance as a training signal. In [37], the authors implement bias mitigation as a way of erasing the sensitive group membership by back-propagating negative gradients in a multi-headed feedforward neural

Model	Accuracy	FPR	FNR	Parity Loss	Pareto Loss
Baseline (no bias loss)	0.879	0.348	0.701	0.192	0.018
Equality Constraint[465]	0.870	0.381	**0.684**	**0.132**	0.123
Adversarial [37]	0.837	0.327	0.723	0.253	0.087
Min-max [275]	0.839	0.306	0.765	0.231	0.055
Pareto Efficient Fair Loss	**0.939**	**0.266**	0.690	0.198	**0.000**

Table 11.2: Comparison of test losses in UCI Heart Disease dataset. PEF optimizes Pareto loss, while [465] minimizes Parity loss. The higher parity loss for PEF does not mean degrading group performances, but instead improves each group. Also, PEF and [465] achieve best False Positive Rate (FPR) and False Negative Rate (FNR) respectively as a side-effect [335], despite not optimizing for it.

network. In [275], they adopt a minimax objective that ensures that the least performing group has the highest accuracy possible. We evaluate by comparing these 4 techniques on the UCI Heart disease dataset. We perform a 10-fold cross validation and report the average accuracy across the 10 splits.

11.4.1 Preprocessing

Each entry in the dataset has been pre-processed using the one-hot encoding for categorical features and the Tensorflow bucketization library into 10 buckets for numeric features. The resulting embedding is concatenated and used as input to a 3-layer feedforward neural network with 256, 128 and 64 hidden units respectively. We trained each of the models for 100 epochs and noticed that training and dev error plateaued. The test metric reported is the average of 10-fold demographic group stratified cross validation accuracy along with the corresponding error bars denoting one standard deviation. The group identifiers present in the datasets were used to aggregate group Pareto loss during training.

11.4.2 Demographic Group Performance

The UCI Heart Disease dataset predicts angiographic disease status as a binary label (presence or absence of Heart Disease) using medical and demographic information. Age is binarized at a threshold of 40 years between young and old individuals, and gender is given to be binary (male/female) and are assigned as sensitive variables. The intersection of these 2 variables are considered sensitive demographic groups in our study. In Figure 11.2, we present group level performances for the UCI Heart Disease Dataset. Our approach of incorporating pareto efficiency leads to improvements in group level accuracies for all groups of the data by an average of 9.6%. We see improvements in the accuracy of predicting the presence of Heart disease in Table 11.2 by an average of 9.7% and that the relaxation of the demographic parity loss performs better than strict fairness constraints (Figure 11.3). This implies that improving based on demographic pareto efficiency obtains a better overall accuracy than even the baseline which explicitly optimizes overall accuracy on a held-out test set. This non-trivial result is due to the fact that when optimizing for overall accuracy on a training dataset, predictive models may incorrectly assume that the patterns in the majority group (Young/Male) might generalize to other demographic groups. We overcome this issue, and ensure that the demographic groups' accuracies are improved in an iterative manner as outlined in Algorithm 1. Since we use the decoupled classifiers' accuracies to measure the Pareto losses, and ensure that we incrementally train the joint model in such a way as to improve the accuracies of each of the groups (the training will terminate if individual group accuracies cannot be improved). This in turn has improved the overall test accuracy by overcoming issues of overfitting to the majority demographic group.

11.4.3 Trade-off Parameters

The choice to optimize overall accuracy as opposed to group-specific pareto efficiency cannot be made blindly. Hence, it is important to understand the impact of λ, α on the group-level accuracy-

fairness trade-off. In Figure 11.4, we do a parameter sweep across values from 0 to 1 in increments of 0.1 and notice the changes in the overall accuracy, and the group-specific accuracies, along with the corresponding Parity and Pareto losses associated with the test evaluation. Based on this grid, the optimal choice of parameters (λ, α) based on overall accuracy is $(0, 0)$, whereas for each of the four demographic groups are $(0.6, 0.1)$, $(0.1, 0.4)$, $(0, 0)$, $(0.3, 0.2)$; whereas the choice for optimizing parity loss is $(0.4, 0.4)$ and the one for pareto loss is $(0.9, 0.5)$. These trade-offs further illustrate the choice required to be made by domain practitioners when adopting a classifier for predicting angiographic disease status. Table 11.2 and Figure 11.2 values are plotted with these parameters into account. We see that in some groups (e.g. Young/Male), the baseline without fairness based bias loss is comparable to a solution that maximizes that group's accuracy. Such baselines although pareto efficient, lie outside the region of relaxation in fairness weight permitted (Fairness Weight = 1-Parity Loss) and are hence not desirable.

11.5 Discussion and Significance

Jointly Trained vs Decoupled Models: The choice of decoupled models in healthcare diagnosis needs to be made with careful consideration of the stratification dimensions. Decoupled models may be applicable when membership in a demographic group has been shown to have clinical significance. If the objective as presented is to maximize individual group level accuracies, one might be tempted to train a model for each strata separately. Our paper demonstrates the need for joint training across demographic strata to achieve pareto efficient fairness. Purely decoupled classifiers are optimal only under certain conditions of distributional uniformity and availability of data [101]. However, our approach works under a real-world skewed data setting where the data for all demographic groups might not be available uniformly, thereby rendering decoupled classifiers to be sub-optimal. When stratified by the chosen set of demographic group attributes, if there is no predictive model in the desired fairness region, our approach performs no worse

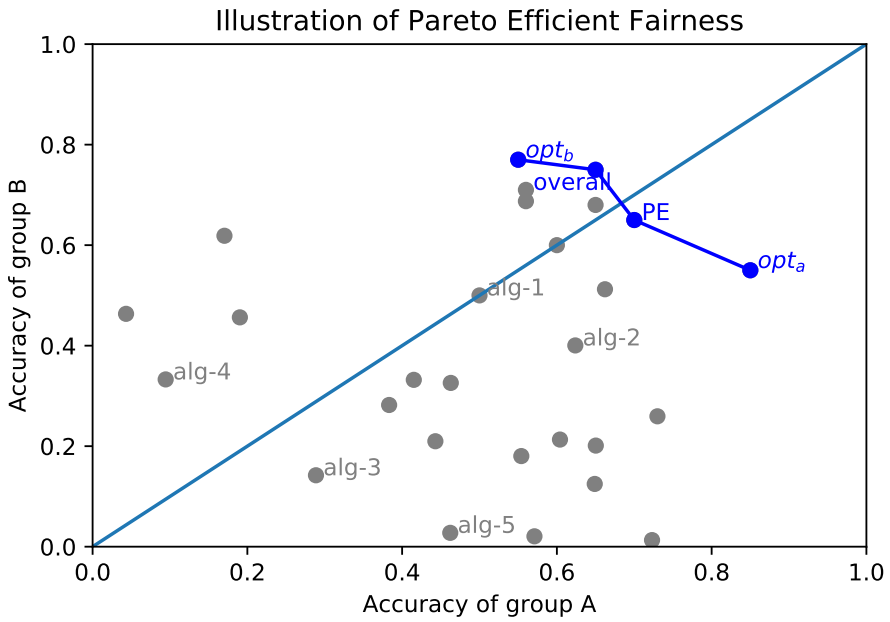


Figure 11.1: Illustration of Demographic Pareto Efficiency on synthetic data. Each point in the scatter plot corresponds to the group level accuracies of machine learning (ML) algorithms (alg-[1-5] indicated in grey) over groups A and B. The best performing ML algorithm with Demographic Parity yields accuracy metrics of $(0.60, 0.60)$ on groups a, b respectively. If accuracy for each of the groups is separately maximized, we would select points $opt_a = (0.83, 0.55)$, and $opt_b = (0.63, 0.77)$. Discovering all the Demographic Pareto Efficient classifiers gives us the Pareto front (dots in blue). Among these Demographic Pareto Efficient classifiers, we could choose $PE = (0.71, 0.63)$ (in blue and green), if our objective was to improve the accuracy metrics of both groups, with minimal deviation from optimal per-group accuracies (pareto loss).

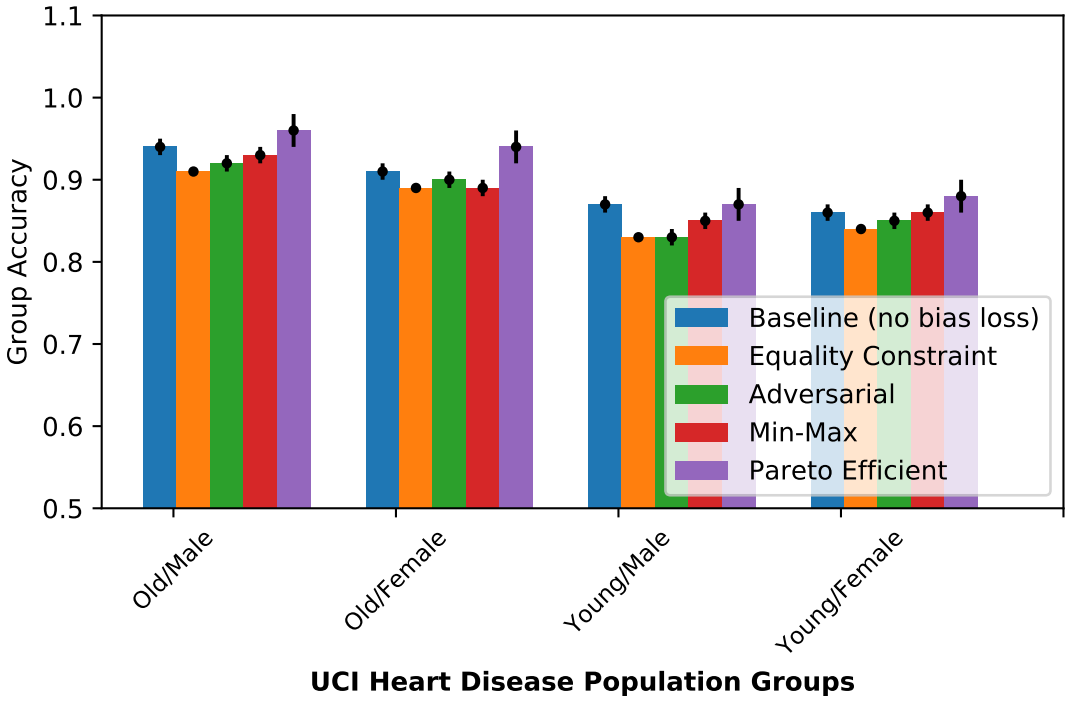


Figure 11.2: Group accuracy comparison showing we achieve Demographic Pareto Efficient group level accuracy for all groups in UCI Heart Disease dataset among constrained classifiers.

than existing equality based constraints as our Pareto loss will be dominated by the high variance in loss between groups. In this scenario, our model would hence chose a low absolute Pareto loss, provided that α, the hyperparameter to trade-off between variance and total value of the Pareto loss is appropriately fine-tuned. Hence, to leverage the benefits of transfer learning, as shown in our evaluation it might be beneficial to bootstrap with decoupled classifiers and train jointly.

Demographic group stratification: The stratification we choose to optimize performance by, depends on what domain experts believe is clinically significant for the disease status diagnoses. For example, age and gender are known to be significant in angiographic disease status in patients, and hence there is a possibility for us to learn different decoupled models. Other possible demographic group stratification can be done based on race and geographical location, as coronary artery disease has been shown to be harder to diagnose in black populations [383],

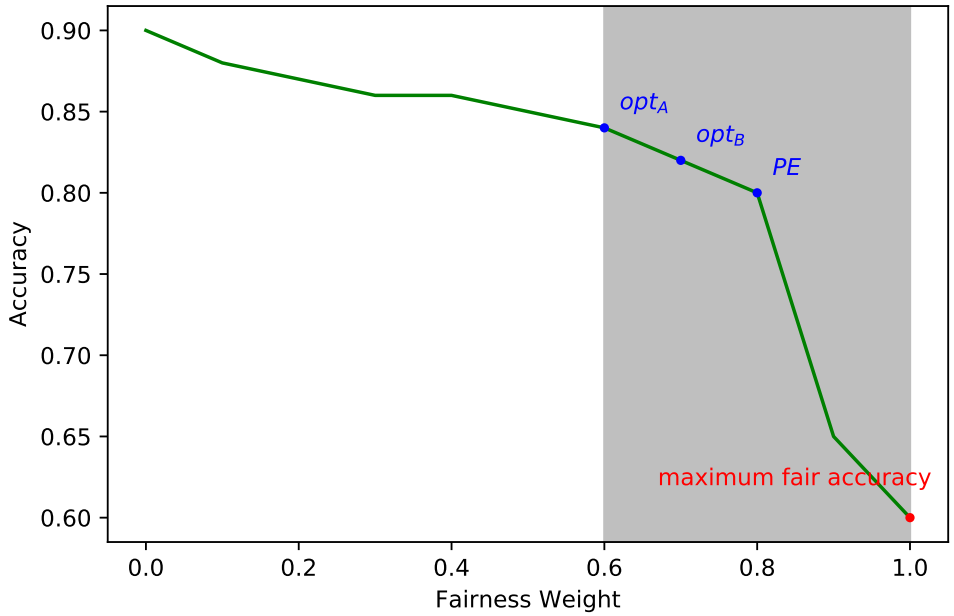
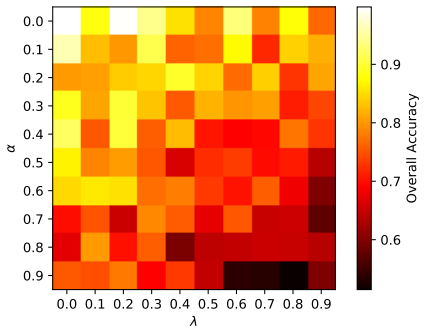


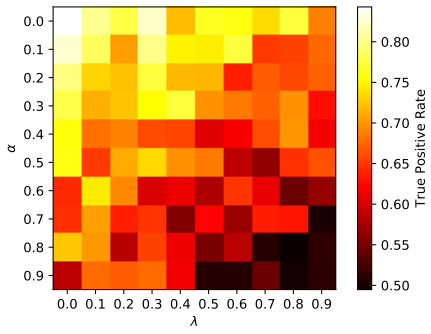
Figure 11.3: Relationship between the shape of the fairness frontier and the efficiency gain expected by using PEF in UCI Heart Disease dataset. y-axis denotes the maximum achievable overall accuracy for a given fairness weight (x-axis). A fairness weight of 1.0 does not permit deviation from the strict equality constraint, wherease a fairness weight of 0.0 is unconstrained and allows higher model performances. However, better accuracies are achievable by relaxing the strict equality constraint by a small amount (gray region) and using PEF.

and that there is a difference in angiographic profiles across patients from different geographical locations in Asia and South America [370, 142, 296, 340].

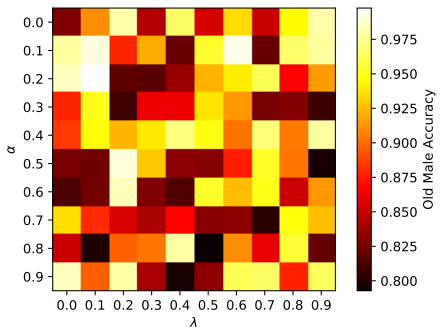
Other definitions of Fairness and Individual Accountability: Notions of pareto efficiency are compatible with assumptions of individual fairness. Utilizing our methodology, the domain expert can make an informed choice among different Pareto efficient models. We have demonstrated that achieving Demographic Pareto Efficiency has benefits and yields classifiers that outperform the baselines for overall and *all* group accuracy. Individual instance based fairness definitions often compare diagnosis and outcomes of one patient with similar patients in a dataset or counterfactual scenarios. However, defining the dimensions of similarity between individuals can be quite challenging for a specific disease type, and should consider the vari-



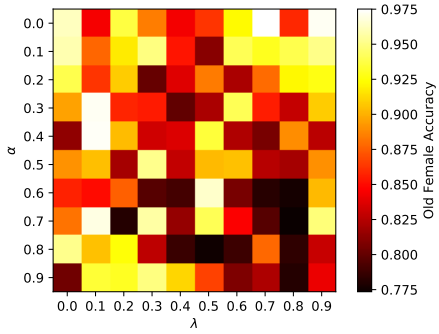
a Overall Accuracy



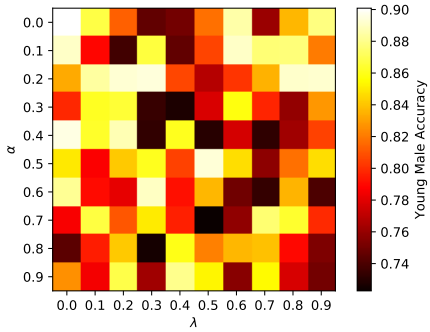
b True Positive Rate



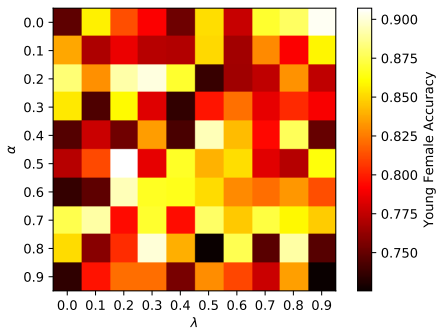
c Old Male Group Accuracy



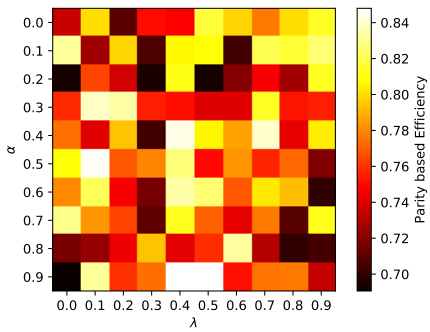
d Old Female Group Accuracy



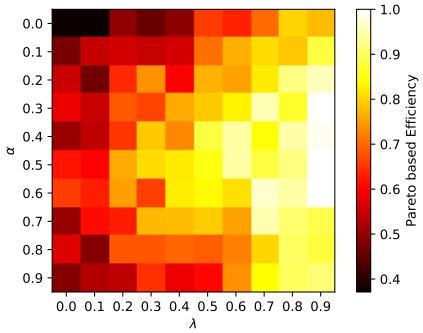
e Young Male Group Accuracy



f Young Female Group Accuracy



g Fairness Weight (1 - Parity Loss)



h Pareto Efficiency (1 - Pareto Loss)

Figure 11.4: Trade-offs between choosing parameters λ and α depends on the group-level versus overall measures chosen by the domain practitioner. Given the prior work that advocates for improving each of the demographic group's accuracy on the Pareto front, we chose our model to optimize Pareto Efficient Fairness (h)

ations of disease prognosis along the same dimensions such as demographic information and co-morbidities. In the event where multiple demographically Pareto Fair operating points are discovered on the Pareto frontier [456], domain experts should choose the right operating point among them by incorporating other procedural steps to mitigate the discriminatory outcomes earlier in the process. Further, pareto efficiency improves the accuracy of minority and under-represented protected demographic groups when compared to unconstrained classifiers, which may implicitly allow the dominance of majority demographic groups when overall accuracy is optimized. While our methodology does not completely eliminate discriminatory biases, Demographic Pareto Efficiency and the choices around it can provide more transparency and understanding of the structural and socio-technical causes behind unfairly distributed datasets and models, which can improve the contestability of ML predictive models.

Other Heart Diseases: In addition to the diagnostic task of angiographic disease status, we see this choice between decoupled and jointly trained models emerge in other heart disease tasks too. In a cardiology study of over 4000 ER patients with cardiac event symptoms [8], no symptoms were found to be predictive of a heart attack in white women. In black males, only an unrelated symptom (diaphoresis) was found to be indicative of a future cardiac event with 95 percent confidence, while in white males, relevant features (left arm radiation, pressure, tightness) were detected as indicators with high accuracy.

11.6 Conclusion

The choice between decoupled and jointly trained diagnostic models for angiographic disease status is critical for positive health outcomes in demographic groups. We have shown that by optimizing for Demographic Pareto Efficiency, the choice between decoupled and jointly trained models can be further broken down to choice of classifiers that have Pareto optimal performance across the demographic groups. As the Pareto front is unknown, we show that by incorporating

a heuristic based on Pareto Efficient Fairness in training a combination of decoupled and jointly trained models, we achieve better overall and individual demographic group level accuracy as compared to other constraints in decoupled and jointly trained models. We demonstrate empirically that our approach achieves Demographic Pareto Efficiency by improving overall and subgroup accuracy by up to 9.7% and 9.6% respectively in the UCI Heart Disease dataset.

12 | Targeted Policy Recommendations using Outcome-aware Clustering

12.1 Introduction

Policymakers and development practitioners aim at implementing policies designed to improve a population's outcomes. However, they often rely on little to no data on what impact the policy recommendations would have at the population level. In the scenarios when observational data is available, econometric models have allowed to determine which input variables have the strongest association with an outcome of interest and have provided guidance on policy recommendations aimed at changing the value of these inputs variables. A fundamental drawback of this approach is that the model would typically prescribe the same set of actions for each individual in a population. In reality, a policy which may appear as the optimal policy on average may not be the best fit at an individual or sub-population-level.

This paper specifically addresses the problem of determining targeted agricultural policy interventions for different sub-groups of the farmer population in Sub-Saharan Africa (SSA) to enhance agricultural outcomes with the ultimate goal of enhancing the livelihoods of the population in the region. The SSA region accounts for more than 950 million people, approximately 13% of the global population. By 2050, this share is projected to increase to almost 22% or 2.1 billion. Agriculture accounts for about 25% of Growth Domestic Product in SSA, and farming is the

primary employment for about 60% of the population. Although that percentage is down from 80% a decade ago, it will remain a major component of economic activity in the SSA region in the coming decade. Given the key role of agriculture will continue to play, it is crucial to design policies aiming at promoting growth and sustainability in that sector.

In this paper, we propose *outcome-aware clustering*, a new methodology to segment a population into clusters that closely match the cluster feature variations with the outcome variations. Given a specific outcome of interest, the primary goal of outcome aware clustering is to segment the population into meaningful and related sub-groups. These clusters provide a framework to the development practitioners on the field, who can then personalize and choose the best outcome-specific predictive policy recommendation and customized support at a cluster-level granularity. This further bridges the gap between the econometric population level modeling, and the practical applicability on the field, where serving the development needs of individual clients is paramount.

Outcome-aware clustering fundamentally differs from the broad array of research on clustering and segmentation. Segmentation of a population, in general, focuses on grouping people into non-overlapping segments such that all the users in the same segment have similar needs and preferences. From a policy perspective, segmentation allows effective customization of policy recommendations to the particular preferences of each segment.

In outcome-aware clustering, the primary objective of clustering is centered on the outcome variable of interest. Conventionally, clustering algorithms have primarily centered around unsupervised learning. The popular k-means (and its variants k-medians, k-medoids, etc.), hierarchical clustering [360], and spectral clustering [381, 305] are notable examples. All these clustering approaches specify a distance/similarity measure between data points and determine the segments by optimizing a merit function that captures the quality of any given clustering. However, the distance function used in these clustering algorithms is independent of any outcome variable.

Outcome-aware clustering performs two key steps to directly tie the outcome variable with

the clustering process. First, given a specific outcome of interest, outcome-aware clustering segments the population based on selecting a small set of features that closely relate with the outcome variable. Outcome-aware clustering measures distance between two users in the population in the reduced feature space. This step essentially makes the clustering process partially supervised. Second, the cluster generation algorithm aims to generate near-homogeneous clusters based on a combination of cluster size-balancing constraints, inter and intra-cluster distances in the reduced feature space.

While outcome-aware clustering normalizes each feature in the reduced space, it specifically does not tie the distance function used in the clustering algorithm to variations in the outcome variable. This is specifically to avoid any specific distance biases that the outcome variable may introduce with respect to specific features in the reduced space. Outcome-aware clustering is also designed for highly noisy contexts where the reduced features may only be weakly correlated with the outcome variable and may only provide limited information about the user with regards to the outcome of interest. Across many survey-based observational studies, especially with missing and noisy entries, we often encounter very few features (sometimes even zero) variables that may exhibit strong correlation with a given outcome variable. Outcome-based clustering is specifically designed to be robust in the face of the observational data having missing values or noisy features or the absence of any features that strongly correlate with an outcome variable.

Outcome-aware clusters can enable field staff to provide customized support based on cluster-level policy recommendations. The basic approach we use to generate targeted policy recommendations for each outcome-aware cluster is a standard multivariate regression based on a condensed set of actionable policy features that are regressed with the outcome variable. These condensed set of variables need to satisfy three properties: (a) Every variable from a policy perspective, needs to be *actionable*, where the policy recommendation is possible on the variable; (b) Every variable should have at least weak correlation with the outcome variable at the cluster level; (c) If a group of two or more variables, exhibit strong co-linearity among themselves, we

reduce these set of variables to the most appropriate variable for the regression analysis.

We demonstrate how the outcome-aware clustering method can be used to the address the problem of improving farmers outcomes in several countries in sub-Saharan Africa (SSA), using data from the World Bank's Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA). Based on a detailed analysis of the LSMS-ISA, we derive outcome-aware clusters of farmer populations across three sub-Saharan African countries and show that the targeted policy recommendations at the cluster level significantly differ from the policies that are generated at the population level. Based on multiple years of LSMS-ISA surveys, we then demonstrate early evidence of movement of populations across clusters for the dominant cluster-specific policy recommendations.

12.2 Related Work

The terms clustering and segmentation have typically been used interchabeably across a broad array of literature spanning multiple disciplines including statistics, machine learning and econometrics. We outline some of the key works that closely relate in spirit to our work. We refer the reader to [450] and [193] for a detailed review of the literature.

The most popular class of clustering algorithms is similarity based clustering, where each algorithm uses a specific distance/similarity measure between data points and determine the segments by optimizing a merit function that captures the "quality" of any given clustering. The popular k-means (and its variants k-medians, k-medoids, etc.), hierarchical clustering [360], and spectral clustering [381, 305] are notable examples. Another class of clustering algorithms is model-based clustering techniques [117, 469] which assume that each cluster is associated with an underlying probabilistic model and different clusters differ on the parameters describing the model. They estimate a finite mixture model [283] to the data and classify customers based on the posterior membership probabilities. However, as mentioned earlier, outcome-aware clus-

tering fundamentally differs from these algorithms in that all these algorithms are completely unsupervised and are not tied to any specific outcome variable or objective.

Outcome-aware clustering also closely relates to customer segmentation literature in operations and statistics. One traditional method for predictive clustering is automatic interaction detection (AID), which splits the population into non-overlapping groups that differ maximally according to a dependent variable, such as purchase behavior, on the basis of a set of independent variables, like socioeconomic and demographic characteristics [20, 264]. [205] proposed hierarchical segmentation techniques tailored to conjoint analysis, which group users such that the accuracy with which preferences/choices are predicted from product attributes or profiles is maximized. Cluster-wise regression methods [435, 436] cluster users in a population such that the regression fit is optimized within each cluster.

Latent class (or mixture) methods offer a statistical approach to the segmentation problem. Mixture regression models [434] simultaneously group subjects into unobserved segments and estimate a regression model within each segment, and were pioneered by [207] who propose a clusterwise logit model to segment households based on brand preferences and price sensitivities. This was extended by [163] who incorporated demographic variables and [206] who incorporated differences in customer choice-making processes, resulting in models that produce identifiable and actionable segments. Existing deep learning based clustering approaches use the dimensionality reduction capabilities of neural networks [181, 446] and learn clustering assignments from the resulting representation [55], but they lack interpretability with respect to the desired outcome. While outcome-aware clustering makes no specific assumptions about the features or the characteristics of the population, many of these latent approaches implicitly assume a mixture distribution characterization that describes the population.

Agriculture policy experts have often relied on demographic attributes to cluster populations [18] or spatial characteristics that also cluster socio-economic indicators like food security, poverty, etc. [11]. However, recent methodologies have advocated for case-by-case analysis

by using decision trees over categorical attributes. Further, policy making frameworks such as FSSIM-Dev (Farming System Simulator for Developing Countries) model the households to understand ex-ante the impact due to roll out of agricultural, technology policies, price changes and market shocks. FSSIM-Dev models each household using mathematical models of risk aversion, capital and crop rotation constraints [256, 175] under assumptions of costs of fertilizers, seeds, water, electricity, etc along with market participation models of individual households using domain-specific polynomial and/or differential equations. The parameters of these mathematical models are then estimated from the observed survey data such as the LSMS-ISA farm household survey which provide priors over parameters, and allow cross-sectional validation and calibration of the model parameters. While our approach shares the common goal of increasing household income, it differs from such a modeling approach by clustering households based on the observed dataset in an outcome-aware manner without access to prior knowledge and provides policy recommendations to improve the outcome that are best supported by the longitudinal survey data tracking farm households over multiple waves of the survey.

12.3 Achieving Agricultural Transformation in Sub-Saharan Africa

12.3.1 Dataset

To understand the factors improving farmers' standards of living, we use data from the LSMS-ISA survey. This survey consists in a nationally representative household panel data with a strong focus on agriculture and rural development. It was designed to improve the understanding of development in the SSA region, in particular of the linkages between farm and non-farm activities.

This survey has been implemented in eight countries in multiple waves. Most of our analysis will focus on the 2015 survey for Ethiopia. In section 12.5, we also show how our results can be

extended to Tanzania and Uganda, comparing our main policy results across countries.

Before delving into the analysis, it is important to understand some of the limitations associated with using the LSMS-ISA dataset to conduct this analysis. First, a significant number of zeros and missing values limits the ability to draw inferences at a subpopulation level. We choose to discard survey answers with more than 30% of missing values. Second, we also drop variables which are not observed across multiple waves.

12.3.2 Relevant Outcomes and Inputs

A policy maker aiming at improving the living conditions of farmers in sub-Saharan Africa could choose to focus on a variety of outcomes: their revenue, level of expenditure, food expenditure diversification, whether they receive medical assistance when they are ill, whether they face food deficiency, etc. We find that among these outcomes of interest, the correlation is only 9% on average (Fig. 12.1a). This suggests that each outcome follows its own path, hence policy recommendations should be independently evaluated for each outcome.

In addition, while a large number of inputs could in principle play a role in farmers' living conditions, inputs with high correlation with outcomes are good candidates to consider when looking to improve farmers' outcomes. For the purpose of deriving policy recommendations, we distinguish between inputs that can be modified through short-term policy actions ("actionable") from those that cannot ("non-actionable").

We find that for inputs with high correlation with outcomes variables, while these correlations typically have the same sign across outcome variables, their magnitude tend to vary substantially (Fig. 12.1b and c). As correlation between outcomes are low, it is not surprising that the effect of a given input will vary across outcomes, reinforcing the conclusion that policy recommendations need to be outcome specific. We also find that even the most impactful input variables only have a 10% correlation with outcome variables on average, leading to a set of less than 10 actionable inputs likely to have an substantial impact on a given outcome.

Figure 12.1: Relationship Between Farmers' Outcomes and Inputs: (a) Spearman correlations between farmers' outcomes, showing a low average correlation equal to 0.09, and suggesting that policy recommendations should be derived for each outcome separately. We also show the Spearman correlations between farmers' outcomes and inputs, separating (b) non-actionable from (c) actionable inputs, and ranking inputs by their average correlation across outcomes. These subplots indicate that for inputs with the high correlations with outcome variables, correlations across outcomes are of similar sign but

12.4 Methodology

Generating policy recommendations can be thought of as a problem of extracting features which are predictive of an outcome intended by the policy. Given a set of n features F in an input variable matrix X, an outcome variable y, we intend to identify the best set of features P which would predict the outcome variable. We now describe our approach in the rest of the section. First, we cluster the features using a novel *outcome-aware clustering* algorithm. We then learn a regression model for each of these clusters separately to identify important actionable variables which significantly predict the outcome variable.

12.4.1 Outcome aware clustering

We define outcome aware clustering as the problem of choosing a subset of features C such that the unsupervised clusters on these features effectively separate both the input features and the outcome variable across these clusters.

Prior to doing any clustering, it is essential to ensure that we don't incorporate features with a large fraction of missing values. Since most features in our study are categorical in nature, using any form of imputation or matrix completion techniques on these would not be sound. Hence, a simple threshold based filtering is used. Normalization of the features used for clustering is done by applying the z-score method.

In addition to finding the features to cluster on, we need to fix on the number of clusters to learn in a commonly used k-means clustering. During each step of making the choices of features to cluster on, we identified k using the elbow method and the average euclidean distance from the centroids across a range of $k \in [1,10]$.

As explained in Algorithm 7, we initialize C as an empty set and iteratively add features from the full set of non-actionable and actionable features (F) to C in a greedy fashion. In each iteration, we choose a feature which maximizes a weighted silhouette coefficient for the k-means clustering

obtained by including the feature in the clustering set C. This weighted silhouette coefficient (sc) combines the sc as measured in the clustering feature space as well as the single dimensional outcome space. The outcome awareness is controlled by a parameter $\alpha \in [0, 1]$. We can see that $\alpha = 0$ is equivalent to traditional unsupervised clustering on the input feature space, whereas $\alpha = 1$ is equivalent to bucketization based only on the outcome variable. With α between 0 and 1, the clustering achieves two objectives. First, we identify a clustering (l_f) which can separate the clusters based on the outcome variable, allowing to design policy recommendations at various outcome levels. Second, it separates the input features space which is critical to identifying these clusters when the outcome variable is not observed in an unsupervised manner.

Algorithm 7 Feature choice for clustering

$F := \{f_1, f_2, f_3, .., f_n\}$, input features
$y :=$ output feature
$\alpha \in [0, 1]$, Output awareness parameter
$C := \emptyset$
$\epsilon :=$ Threshold of k-means silhouette coefficient (sc) improvement
while $\Delta sc > \epsilon$ **do**
  for f in $F \backslash C$ **do**
    $l_f = Kmeans(f \cup C)$
    $sc_{y,f\cup C} = \alpha * sc_y(l_f) + (1 - \alpha) * sc_{f\cup C}(l_f)$
  end for
  $f_{opt} = \underset{f \in F\backslash C}{\text{argmax}}\ sc_{y,f\cup C}$
  $\Delta sc = sc_{y,f_{opt}\cup C} - sc_{y,C}$
  $C := f_{opt} \cup C$
end while
return C

A benefit of choosing the features iteratively is that we don't end up with redundant features which explain the same feature space and outcome level. This ensures that the final set of features can distinguish between any pair of clusters using only a subset of these features. This can be thought of increasing the information criterion of the clusters iteratively. Hence, some of the features chosen during the iterative steps could have low outcome correlation values at the population level, but are instrumental in distinguishing certain specific outcome clusters. In each

step, the k-means also enforces that each cluster is of a certain minimum size to avoid learning behavior of statistical outliers, and guarantee that we have enough observation to derive cluster-level policy recommendations.

The stopping condition of iterations is based on the improvement in the silhouette coefficient over the iterations, and the threshold (ϵ) can be chosen in a problem specific manner. Once the feature set C is chosen, we have also jointly learnt the corresponding k-means clusters. It can be noted that our algorithm is generic and can accommodate any unsupervised clustering method and operates as a layer above it.

12.4.2 Policy recommendations through regression

The fundamental contribution of our approach is that we learn different policy recommendations for different clusters of households. These variations in policy recommendations across clusters are not evident if done at a population level.

As shown in Algorithm 8, choosing features for regression is done in a principled two step approach. First, we used highly correlated features with the outcome, where a threshold (β) on the spearman correlation coefficient (ρ) was used for filtering. Second, in order to eliminate multi-collinearity in the correlated features, we iteratively eliminated the feature with the highest variance inflation factor (VIF) above a certain threshold (γ). These thresholds were identified using an appropriate grid search to ensure that a reasonable set of policy recommendations were identified. The filtered features are then used in a linear regression model to predict the outcome variable for each cluster. Statistically significant coefficients of this model are then used to derive policy recommendations for each cluster.

Algorithm 8 Regression based Policy Recommendations

$C := \{c_1, c_2, .., c_k\}$, the set of clusters
$F := \{f_1, f_2, f_3, .., f_n\}$, set of actionable features
$y := (obs, 1)$ output matrix
$X := (obs, n)$ input matrix
$\beta :=$ Output correlation threshold
$\gamma :=$ Input multi-collinearity threshold
$F_{corr} = \{f_i | \rho(y, X[f_i]) > \beta\}$
repeat
  $f_{max} = \underset{f \in F_{corr}}{\text{argmax}}\ VIF(X[F_{corr}], X[f])$
  $F_{corr}.remove(f_{max})$
until $(VIF(X[F_{corr}], X[f_{max}]) < \gamma)$
for c in C **do**
  $coeff_c = OLS(X_c[F_{corr}], y_c)$
  $P_c :=$ stat-significant $coeff_c$
end for
return $\underset{c \in C}{\cup} P_c$

12.5 Results

12.5.1 Clustering Farm Households

Next, we experiment the clustering method that we have developed on the 2015 LSMS-ISA survey of Ethiopia. We focused on farmers' crop sales as our outcome of interest. Our algorithm suggested to cluster farm households based on the following inputs: their total land surface, household size, the number of oxen they own, the number of ploughs they own, whether or not they participate in an extension program, the quantity of chemical fertilizers they use, and their number of hired workers. These inputs are indeed among those having the highest correlation with crop sales. We then allocate households into four clusters as suggested by the Elbow method (Fig. 12.2c).

We find that our clustering method indeed allows to construct clusters in which households crop sales are similar within each cluster and different across clusters (Fig. 12.2a). On average,

crop sales increases monotonically across clusters, ranging from 711 Birr to 2,424 Birr. Projecting our clustering inputs on their first two principal components, we also find that our method allows to construct clusters in which clustering inputs are similar within each cluster and different across clusters (Fig. 12.2b).

Compared to all of the richer clusters, households in the first cluster only own 0.24 Ha of land on average, which is 6.8 times less than households in the second cluster (Fig. 12.2d). They are comprised of five members on average, compared to six for the other clusters (Fig. 12.2e). They are five times less likely to own an ox (Fig. 12.2f) and 2.4 times less likely to own a plough (Table 12.1) compared to households in the second clusters. 28% of them are female-headed households (Table 12.1), which is 1.8 times more than in the other clusters, and they are predominantly located in the SNNP region (Fig. 12.3). These are the poorest households in our sample; they do not have the means to own large properties nor the ability to purchase basic tools required to harvest efficiently.

Households in the second clusters generate 1.8 times more revenue and are better equipped than those of the first cluster. Yet, they still do not use significant amounts of fertilizers (Fig. 12.2g) or improved seeds (Table 12.1) to increase their productivity compared to those in the third or fourth cluster. Only about 13% of households in the first two clusters participate in an extension program, and only about 13% of them use damaged prevention techniques, compared to about respectively 76% and 22% of those in the last two clusters (Fig. 12.2h and Table 12.1). Only about 12% of households in the first two clusters use credit services, compared to about 27% of those in the third or the fourth cluster (Fig. 12.2i).

The richest households are located in the fourth cluster, with a average income 60% larger than those of the third cluster. They are mainly characterized by their ability to hire workers (Fig. 12.2j). 22% of them save money, compared to less than 15% of households in third clusters and below. They also tend to acquire more sophisticated or more expensive tools. They are 2.1 times more likely to own a pick ax (Table 12.1), and 1.5 times more likely to own an ax (Table

12.1) compared to those in the third cluster or below.

Taken together, these results show that the clusters derived from our *outcome-aware clustering* are robust and correspond to interpretable subpopulations of households.

12.5.2 Policy Recommendations

Having constructed robust and interpretable clusters, we now ask whether we can derive policy recommendations at the cluster level, and whether these recommendations differ from those obtained at the population level.

As our analysis is conducted on a relatively small dataset, we choose to estimate a multivariate regression model of crop sales using a restricted set of policy variables. We apply algorithm 8 choosing the two following parameters: (a) we remove any policy variable that has a correlation with crop sales of less than $\beta = 0.05$, and (b) we iteratively remove policy variables until the VIF scores of the remaining variables is less than $\gamma = 1.5$. This guarantees that the selected variables will have a substantial impact on the outcomes, and will remove collinear policy variables from the model. In a robustness check, we found that our results hold for a wide range of values for β and γ, other specifications typically leading to a larger set of insignificant variables being included in the model.

The number of hired workers has the strongest coefficient in the full sample regression (Fig. 12.4a). As the standard deviation of crop sales is equal to 1,169 Birr, hiring one additional worker is associated with an increase in income of $0.25 \times 1,169 = 292$ Birr. The effect of hiring workers on crop sales is U-shaped, with the largest effect concentrated in the first cluster where the coefficient is equal to 0.7. It indicates that policies should primarily focus on encouraging farmers to hire workers, especially in the first and the fourth cluster. Possible implementations could be to subsidize workers hiring costs, develop or improve systems providing information on labor market conditions, etc. It is important to note that our analysis does not account for the costs of implementing such policies. Hiring workers could be quite costly, especially for low income

households. []

The second most impactful factor corresponds to the use irrigation techniques (Fig. 12.4b). Households using irrigation have an average revenue that is 128 Birr higher than those who do not. Here, the effect is also U-shaped: it is positive and significant for households in the first and the fourth clusters, but it is insignificant for those in the second and third cluster.

An increase in the quantity of chemical fertilizers used by one standard deviation or in the number of axes owned by one unit are associated with a small increase in income of 105 Birr and 47 Birr respectively (Fig. 12.4c and f). This effect is concentrated on households in the first cluster, the effect being insignificant for the remaining clusters. This suggests that policy aiming at improving the income prospects of households in the first and second clusters specifically could be targeted towards reducing the costs of acquiring additional tools or fertilizers through subsidies or conditional cash transfers.

Finally, households using damage prevention techniques or saving money generate on average 94 Birr and 82 Birr respectively more than those who do not (Fig. 12.4d and e). The effect is concentrated on households in the third and fourth cluster and is insignificant for households in the first and second cluster. This suggests that policies targeted towards the third or the fourth cluster could focus on raising awareness on the benefits of damage prevention techniques, or incentivize farmers to save money using their mobile phone.

Taken together, these results show that *outcome-aware clustering* allowed us to derive policy recommendations at the cluster level, showing that they often differ from those that would be optimal at the population level.

12.5.3 Cross-country Comparison

Next, we compare the results that we obtained in Ethiopia to other countries included in the LSMS-ISA survey. We apply outcome-aware clustering on the 2014 survey for Tanzania and the 2013 survey for Uganda, deriving policy recommendations at the cluster level. Although

cross-country comparisons are limited by a lack of homogeneity in how key policy variables are measured across countries, it is nonetheless interesting to test whether some consistent patterns emerge.

The amount of pesticides used has the strongest association with crop sales, both for Tanzania and for Ethiopia. In both cases, the effect is slightly decreasing across clusters (Fig. 12.4g and l).

The next variable with the strongest association with crop sales both for Tanzania and for Uganda is the amount of fertilizers used (Fig. 12.4h and l). The strength of the effect is U-shaped across cluster for Tanzania, and has an inverted U-shaped for Uganda, which differ from the pattern observed for Ethiopia. These differences could be explained by variations in the variety of crops that are being grown, the relative returns to using fertilizers, or the types of fertilizers being used.

For Tanzania, owning a plough has an effect on crop sales that is mostly concentrated in the first cluster (Fig. 12.4i). This is consistent with the effect of owning an axe being concentrated in the first cluster in the case of Ethiopia.

In the case of Uganda, the effect of hiring workers is not as predominant as in the case of Ethiopia (Fig. 12.4m), yet we observe a similar U-shape behavior.

Finally, having a bank account in Tanzania is only associated with generating more revenue for households in the third and fourth cluster, which is similar to the effect of saving observed for Ethiopia. Similarly, borrowing is associated with a reduction in income only for households in the fourth cluster in the case of Uganda.

12.5.4 Validating Predictions Over Time

To validate our policy recommendations, we do a longitudinal evaluation tracking households across 3 waves of surveys done in Ethiopia, with a gap of 2 years between each wave.

For a majority of households, the value of key inputs remain constant between surveys, limiting the ability to test the validity of our predictions over time. We focused on households'

“number of hired workers", as it is most impactful input coming out of the model predictions, the other inputs being associated with insignificant evidence of movement between waves.

We found evidence of a lift in the increase crop sales associated with hiring an additional worker being equal to 0.39, 0.23, 0.26 and 0.57 across clusters of increasing income (Fig 12.5). This indicates that households in the first cluster who hired an additional worker between two consecutive wave are 39% more likely to have had an increase in crop sales during the same period that those who did not, similar conclusion being drawn for the other clusters. Interestingly, we find a U-shape in the value of the lift factor associated with hiring an additional worker, which mimics the variations in coefficient strengths obtained in the multivariate regression. Although additional data would be needed to provide further evidence, this gives some initial validation for our approach. The average lift evidenced from unsupervised clustering ($\alpha = 0$) and bucketed crop sale buckets ($\alpha = 1$) are 0.21 and 0.16 across 4 clusters respectively, as compared to the average overall longitudinal lift of 0.36 (from Fig 5) we observed over 3 waves of surveys in Ethiopia.

12.6 Discussion

12.6.1 Domain Knowledge based Clustering

Our choice of features to cluster works under the assumption that the policy variables can be intervened upon directly rather than through unobserved underlying factors not in the model. We acknowledge that there might be several such factors in the world, and it might be untestable to validate that no such factor exists in the wild [329]. Compared to other policy targeting procedures based on domain expertise, we provide a framework to group samples into clusters and select policy factors in such a way that attempts to maximize correlation over a set of factors over clusters in the observed dataset. Further, our framework can be used to augment existing policies based on domain expertise, by using the domain-expertise based segmentation features

as an initial set of selected features (C) and then searching for any additional features based on Algorithm 1. This would further build upon the domain expertise and leverage the outcome-feature correlations to optimize clustering. If the domain knowledge-based clustering is robust enough and no new additional features improve the silhouette coefficient of the clustering, then our framework would terminate and provide empirical validation of the choice of the features provided by the domain expert.

However, the chosen feature set might lead to misinterpretation when additional domain knowledge of underlying factors exists. For example, our policy recommendation of improving hired labor, might be controlled by an underlying factor such as overall household wealth for which we do not have a good estimate from the survey. Further, the interpretation of how these policy variables can be intervened upon depends on contextual information such as cost, operability, infrastructure issues that impact the intervention, which goes beyond the scope of what was observed in the dataset. Given such contextual constraints, it is plausible that the various policy recommendations we provide might translate to a few actionable initiatives such as cash vouchers or credit programs that can be operationalized in the field [6].

12.6.2 Evidence based Policymaking

Randomized control trials and contextual bandits have produced desirable and robust treatment assignments, and allow policymakers to accurately estimate the impact of intervening on specific policy variables [27, 89, 432]. However, they incur the cost of experimentation which may be infeasible if we are given access to an observational survey dataset alone. We agree that such methods are superior, but absent the additional cost of setting up a large experimental study across thousands of households in fragile countries, our approach provides additional guidance based on observational data for policymakers in choosing outcome-aware clusters over survey data. Since our method is focused on providing policy guidance under observational data settings without the identification of control and treatment groups, we report the lift observed in

crop sales when the corresponding input features which were identified as possible policy interventions also increased in subsequent waves of survey data in Figure 5. Based on such evidence, the policymaker may then choose to invest in an experimental study to confirm the impact of an actual intervention through randomized control trials in the field. This step may be useful to further eliminate confounding factors that might impact the generalizability and validity of the policy recommendations.

12.6.3 Ethical Considerations

As compared to an experimental study where data is gathered with a known outcome in mind, using observational studies to derive policy recommendations can raise concerns of misuse of the data gathered, especially if used to derive private information of households. For these concerns, we refer to the data management plan in the LSMS-ISA framework [1] which confirms that "*For purposes of maintaining the confidentiality of the data all names and addresses including contact addresses and field descriptions in the post planting agriculture questionnaire have been removed from the datasets. In addition, the GPS coordinates have also been removed as these could be used to locate households and fields with accuracy.*". Further, using the data gathered for improving outcomes of the farmers might lead to inconsistencies in how the data collecting questionnaire was presented to the household members. For example, we assume that the propensity of participating and providing information in an informational survey is the same as when certain incentive structure is associated with the survey, which needs to be further validated by documenting the purpose of data collection for all features. [133]

[1] https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA

12.7 Conclusions

This paper presents *outcome-aware clustering*, a new clustering methodology to segment a population into meaningful clusters corresponding to a specific outcome of interest. Unlike traditional unsupersived clustering and mixture modeling approaches for population segmentation, *outcome-aware clustering* relies on choosing a set of clustering features closely related to an outcome of interest, while minimizing intra-cluster and maximizing inter-cluster distances. We demonstrate the utility of this *outcome-aware clustering* methodology to enable field practitioners to provide personalized and customized cluster-level policy recommendations. Using data from the LSMS-ISA survey across three countries in Sub-Saharan Africa, we found that our method provides actionable and highly predictive cluster-level policy recommendations which significantly differ from those obtained at the population level.

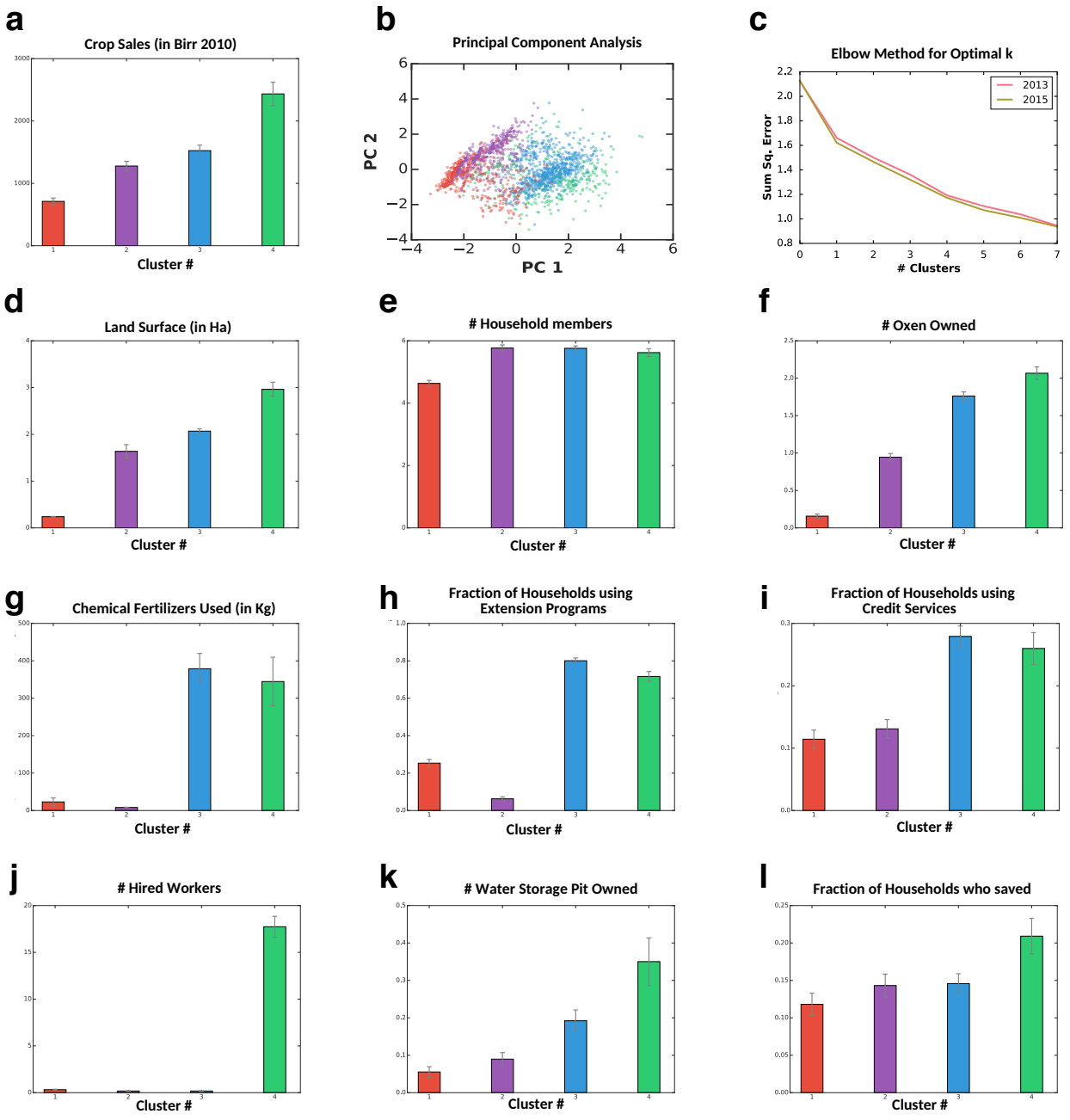


Figure 12.2: Clustering Results: (a) Average crop sales across clusters, indicating that our method allows to construct clusters such that households outcomes are similar within each cluster and different across clusters. (b) The two principal components of our clustering features across households, indicating that our method allows to construct clusters such that households clustering inputs are similar within each cluster and different across clusters. (c) Sum of square errors of K-means clustering, showing that the error is stable across survey waves. The elbow method indicates that the optimal number of clusters is 4. To understand the composition of the resulting clusters, we then show the average value across clusters of the three features with the highest relative change occurring between cluster one and two (d-f), between cluster two and three (g-i), and between cluster three and four (j-k).

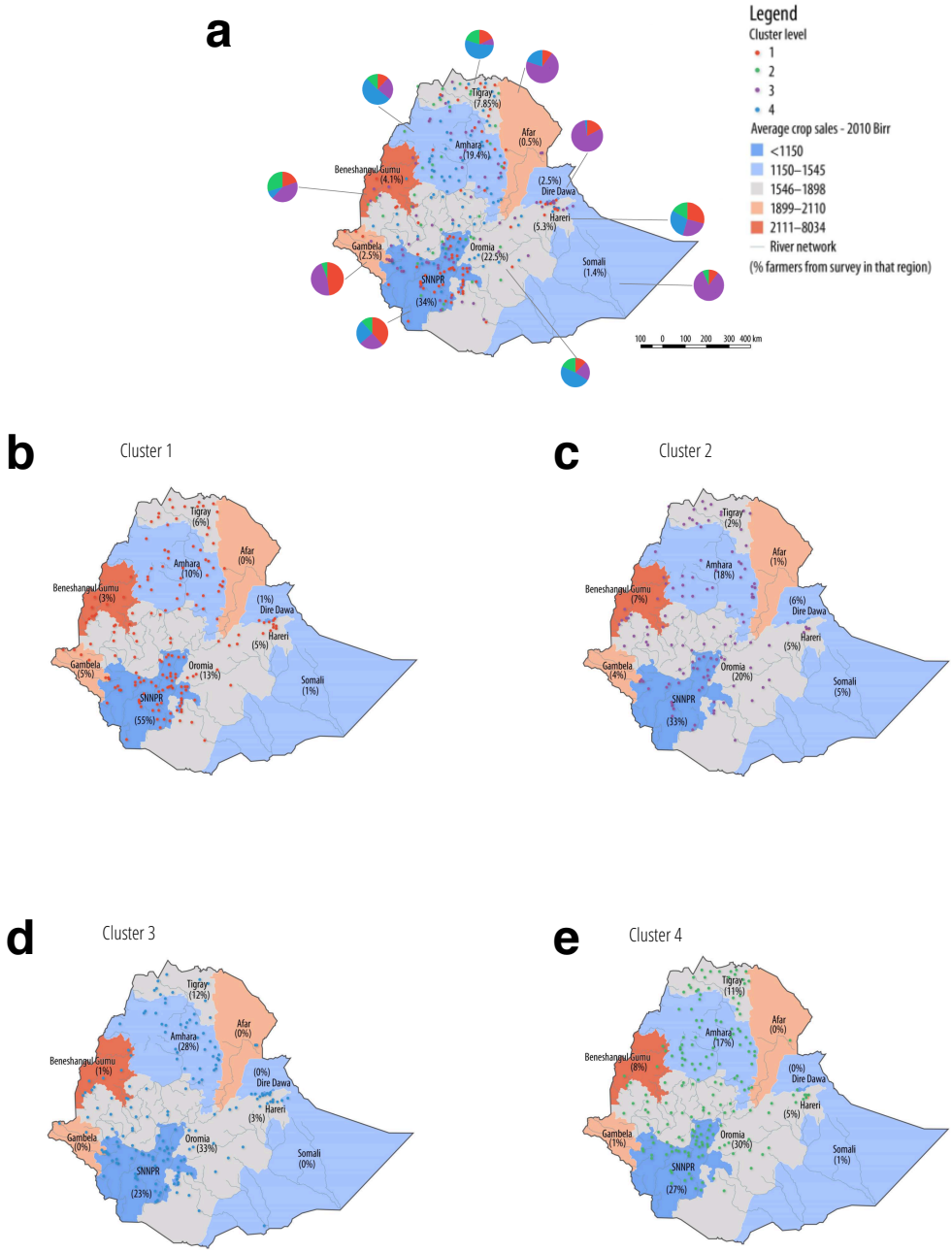


Figure 12.3: Geography of Clusters: Each dot corresponds to a household colored by its cluster.

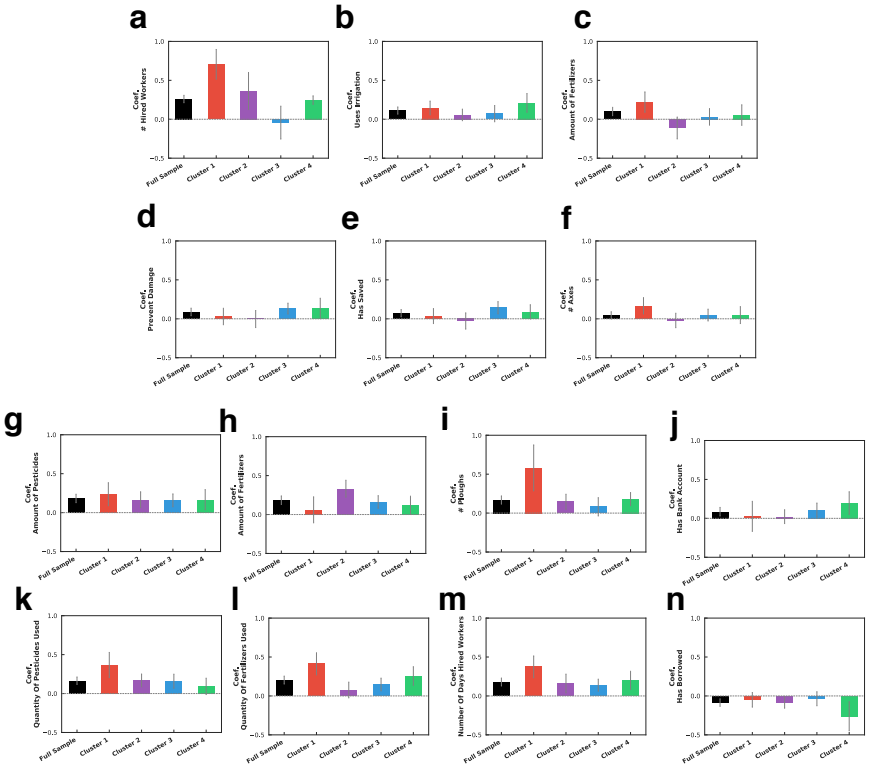


Figure 12.4: Policy Recommendations: Regression coefficients of a multivariate regression of crop sales on a set of selected policy variables, for the entire sample (black), and per cluster of increasing crop sales. Coefficients are ranked by decreasing value on the entire sample. The first two rows corresponds to the 2015 survey for Ethiopia, the third row corresponds to the 2014 survey for Tanzania, and the fourth row corresponds to the 2013 survey for Uganda. This plot shows that the effect of the most impactful variables vary significantly across clusters, indicating that policy recommendations should indeed be cluster-specific.

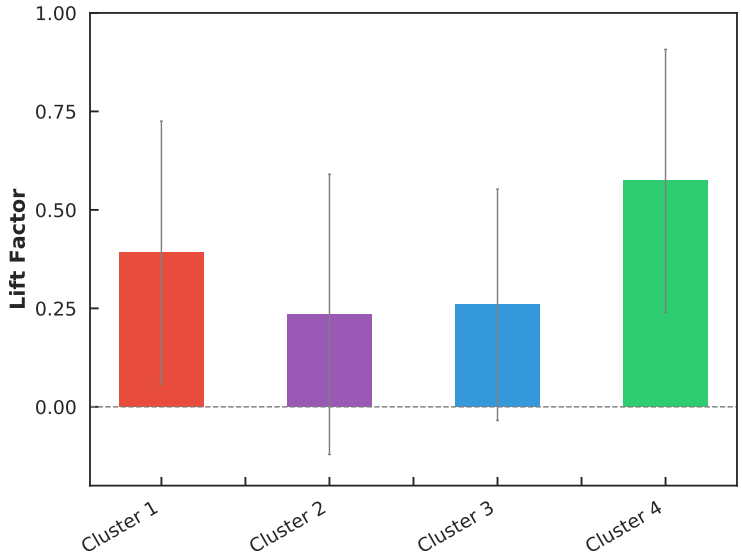


Figure 12.5: Evidence of Movement Between Clusters: For each cluster, the lift factor associated with a given input measures the fraction of households whose income increases beyond a given threshold during two consecutive survey wave when the value of that input also increased, relative the fraction of households whose crop sales increased beyond the same threshold. We pick the threshold to correspond to the 25%ile of the distribution of changes in crop sales for each cluster and each wave. We only show the lift associated with hiring additional workers, the lift associated with less impactful policy inputs being insignificant.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Avg.	Stdev.	Avg.	Stdev.	Avg.	Stdev.	Avg.	Stdev.
Amount Of Assistance Received	51.036	228.505	84.745	356.512	51.177	248.418	57.666	381.920
Attended School	0.320	0.389	0.295	0.384	0.290	0.379	0.363	0.427
Average Precipitation	1271.081	277.231	1210.551	354.563	1228.689	303.100	1260.963	326.223
Average Temperature	181.735	24.328	190.300	32.059	175.884	25.541	192.661	30.492
Children Education	0.696	0.331	0.700	0.343	0.740	0.310	0.719	0.326
Number of Crops Planted	3.638	2.642	2.948	3.184	2.831	3.332	2.778	3.754
Crop Sales (in Birr 2010)	711.440	1170.191	1277.643	1768.510	1524.358	2373.700	2427.224	3195.684
Distance To Market	63.565	42.324	72.449	48.111	60.292	42.312	67.126	46.472
Distance To Population Center	27.208	20.145	40.966	26.594	32.082	19.888	40.130	32.136
Distance To Road	11.713	12.511	17.654	20.798	12.230	11.732	11.940	13.820
Elevation	1998.337	411.404	1910.413	501.948	2138.493	412.222	1850.324	472.553
Non-food Expenditure (in Birr 2010)	1065.626	1460.336	1231.502	1287.209	1775.106	1358.489	2397.284	2195.733
Food Expenditure Diversification	0.840	0.169	0.846	0.147	0.871	0.120	0.875	0.106
Fraction of Households With A Bank Account	0.045	0.207	0.020	0.142	0.049	0.216	0.077	0.267
Has Borrowed	0.227	0.419	0.271	0.444	0.309	0.462	0.253	0.435
Fraction of Households Using Medical Assistance	0.198	0.291	0.210	0.249	0.231	0.268	0.272	0.273
Fraction of Households Who Saved	0.116	0.320	0.144	0.351	0.146	0.354	0.208	0.406
Heavy Rains Preventing Work	0.041	0.219	0.035	0.199	0.039	0.252	0.031	0.413
Household Head Age	47.068	16.779	48.417	15.360	47.426	13.861	46.572	13.659
Fraction of Divorced	0.073	0.261	0.030	0.168	0.016	0.123	0.005	0.072
Fraction of Female-headed Households	0.278	0.448	0.094	0.292	0.147	0.354	0.114	0.318
Fraction of Male-headed Households	0.722	0.448	0.906	0.292	0.853	0.354	0.886	0.318
Household Head Is Monogamous	0.718	0.450	0.854	0.351	0.845	0.361	0.825	0.379
Household Head Is Polygamous	0.024	0.152	0.038	0.191	0.023	0.149	0.067	0.250
Household Head Is Separated	0.003	0.057	0.002	0.042	0.002	0.047	0.008	0.091
Fraction of Widow	0.174	0.379	0.076	0.263	0.103	0.303	0.091	0.287
Household Head Never Married	0.007	0.085	0.000	0.020	0.012	0.109	0.003	0.053
Number of Household Members	4.637	2.087	5.773	2.196	5.754	2.085	5.621	2.158
Illness Of Household Member	0.300	1.032	0.389	1.223	0.294	0.817	0.345	0.864
Increase In Price Of Inputs	0.172	0.474	0.172	0.514	0.265	0.519	0.363	0.553
Land Surface (in Ha)	0.239	0.144	1.638	3.249	2.066	1.400	2.953	2.595
Latitude	7.879	2.021	9.057	2.320	9.361	1.880	9.076	2.058
Literacy Rate	0.325	0.381	0.318	0.369	0.335	0.372	0.405	0.392
Lives In Afar	0.000	0.015	0.001	0.036	0.000	0.015	0.000	0.000
Lives In Amhara	0.126	0.332	0.303	0.460	0.310	0.462	0.235	0.424
Lives In Benishangul Gumuz	0.014	0.119	0.030	0.171	0.004	0.064	0.035	0.183
Lives In Dire Dawa	0.001	0.038	0.007	0.086	0.000	0.009	0.000	0.000
Lives In Gambella	0.006	0.076	0.009	0.095	0.000	0.000	0.001	0.036
Lives In Harari	0.002	0.047	0.002	0.048	0.001	0.033	0.003	0.052
Fraction of Households Living in Oromiya	0.169	0.374	0.352	0.478	0.517	0.500	0.507	0.500
Lives In Snnp	0.650	0.477	0.266	0.442	0.121	0.327	0.161	0.367
Lives In Somalie	0.001	0.025	0.017	0.129	0.000	0.000	0.005	0.069
Lives In Tigray	0.031	0.173	0.011	0.106	0.046	0.210	0.054	0.225
Longitude	38.128	1.198	38.102	1.807	38.190	1.411	37.767	1.521
Fraction of Households Without Food Deficiencies	0.466	0.499	0.689	0.463	0.773	0.419	0.836	0.368
Number Of Axe Owned	0.651	0.695	0.682	0.851	0.545	0.848	0.888	1.065
Number Of Droughts	0.283	0.604	0.434	1.134	0.207	0.567	0.259	0.509
Number Of Hired Workers	0.317	1.075	0.170	0.589	0.177	0.574	17.680	19.054
Number Of Oxen Owned	0.157	0.648	0.950	1.124	1.759	1.449	2.058	1.461
Number Of Pick Axe Owned	0.581	0.720	0.776	0.861	0.831	1.121	1.715	4.761
Number Of Plough Owned	0.315	0.540	0.770	0.634	1.220	0.885	1.239	1.046
Number Of Sickle Owned	1.016	1.011	1.576	1.325	2.155	1.703	2.067	1.766
Number Of Water Storage Pit Owned	0.055	0.306	0.090	0.395	0.192	0.773	0.349	1.081
Fraction of Households Who Own A Land Certificate	0.429	0.486	0.541	0.478	0.665	0.443	0.622	0.447
Percentage Of Damaged Crop	12.551	16.531	21.273	23.839	17.784	20.482	17.693	19.463
Prevent Damage	0.133	0.310	0.124	0.241	0.236	0.288	0.205	0.264
Price Rise Of Food Item	0.304	1.204	0.372	1.365	0.155	0.446	0.158	0.610
Yield (in BIRR per Acre)	5626.935	29955.749	1264.107	1960.196	859.062	1066.618	1278.399	2122.692
Quantity Of Chemical Fertilizers Used (in Kg)	22.925	229.296	7.733	19.063	378.077	1093.361	343.620	1103.675
Quantity Of Improved Seeds Used (In Kg)	2.104	4.641	0.916	5.102	11.835	48.431	12.767	54.739
Rooting Conditions : Mainly Non-Soil	0.003	0.056	0.004	0.064	0.004	0.065	0.000	0.013
Rooting Conditions : Moderate Constraint	0.324	0.468	0.138	0.345	0.184	0.388	0.193	0.395
Rooting Conditions : No Or Slight Constraint	0.466	0.499	0.503	0.500	0.541	0.498	0.618	0.486
Rooting Conditions : Severe Constraint	0.084	0.278	0.202	0.401	0.146	0.353	0.059	0.236
Rooting Conditions : Very Severe Constraint	0.123	0.329	0.153	0.360	0.125	0.330	0.130	0.337
Rural Household	0.960	0.197	0.970	0.171	0.997	0.053	0.985	0.120
Fraction of Households Using Credit Services	0.112	0.315	0.131	0.336	0.280	0.444	0.259	0.434
Fraction of Households Using Extension Programs	0.251	0.433	0.063	0.242	0.800	0.392	0.714	0.448
Uses Irrigation	0.025	0.136	0.029	0.142	0.027	0.105	0.026	0.099
Variations In Greenness	45.215	7.021	45.538	10.094	48.546	8.266	48.560	9.903

Table 12.1: Clusters' Descriptive Statistics

13 Specification Framework for Domain Faithful Deep Learning Systems

13.1 Introduction

Black box deep learning models trained on only observed historical data can make costly errors, which limit their widespread deployment in scenarios that require domain knowledge [310, 385, 448]. Domain experts in these scenarios are particularly skeptical as black-box machine learning (ML) models often contradict rules derived from domain knowledge that has been validated through intervention-based studies like randomized control trials. Even if a model is accurate on historical data, not making use of domain knowledge can limit usefulness [405]. Hence, we propose to build **Domain Faithful Deep Learning** systems, that translate expert-understandable domain knowledge and constraints to be faithfully incorporated into learning deep learning models. In high-stakes domains like health, socio-economic inference and content moderation, a fundamental roadblock for developing deep learning systems is that machine learning models' predictions diverge from established causal domain knowledge when deployed in the real world and fail to faithfully incorporate domain specific structure in counterfactual data distributions. Our framing of domain faithfulness builds on existing robustness research, and aims to build these

ML systems by collaborating with domain experts and addressing critical research questions such as "What data distributions do domain practitioners care about?", "How to faithfully convert domain knowledge into formal constraints expressing the safety, correctness, and fairness of deep learning models?", "How to exploit such model constraints for better generalization?" and finally "How to formally verify that the ML models we learn are grounded in the domain knowledge and in what ways do they deviate?".

In this paper, we provide a framework for domain experts to specify their domain knowledge in a way it can be directly incorporated into training deep learning models. By doing so, we enable the domain experts to answer the above questions by providing implementations of the loss functions through approximations of strict constraints so that they can be used in an automatic gradient (autograd) and compilation optimization framework like JAX [122]. Prior work in these efforts have focused on specifying priors over models using a family of functions, and fine-tuning the function's parameters. Attempts to learn soft-labels through labeling functions are also hindered the roadblock of explicitly writing the labeling functions, which domain experts are not quite familiar with. Hence, a no-code or low-code framework which can provide the domain experts with the necessary specification capabilities which can then be translated into implementations under-the-hood are necessary for ease of adoption. Further, we implement evaluation modules which provide feedback to the domain experts in fine-tuning the parameters of their specifications towards a more robust and domain faithful deep learning system. Our framework, when evaluated on synthetic discontinuous and non-differential constraints perform equivalent to Bayesian modeling techniques [401], while outperforming Bayesian approximation methods on a real MIMIC-III medication recommendation task by 19.1% and a corresponding manually augmented and regularized baseline by 5.3% in Area under the precision-recall graph. We also provide instances of several commonly used constraints, augmentation modules, and also provide an evaluation framework which has shown to be instrumental in achieving the 5.3% improvement in performance by continuously tuning the parameters of the constraints and aug-

mented data distributions.

13.2 Motivation

Since the adoption of the Occam's razor principle, domain knowledge has been known to significantly impact the specifications of a machine learning model, with less features and model capacity required if we have a good prior that positively influences domain generalization. However, arriving at the right domain knowledge to incorporate can be tricky even for the domain experts working with machine learning. For example, consider the medication recommendation task where doctors use patient symptoms as input and recommend a medication as output. Here, although medical ontologies provide guidance on how medications should be prescribed conditional on the patient symptoms, but are not considered to be exhaustive. In fact, doctors use several unstructured sources of priors that are not effectively captured in medical ontologies directly that influence their prescriptions. Hence, it is important to use the medical ontologies as priors when it's available, and rely on data driven patterns when they are not. However, the dichotomy is not trivial as within a patient there might be symptoms which are covered by the ontologies, and some which are not - leading to interactions between these sub-spaces and tackling them requires another level of probabilistic modeling. Hence, whether to incorporate ontologies as priors of models by constraining to a certain family of functions or by adding a regularization penalty over counterfactual augmented data is a choice that domain experts make based on their confidence in those priors. The ability to seamlessly shift between these two choices for instance, is not trivial from an ML engineering persepctive as the current state of the art requires adopting two different modeling perspectives. For the first choice, we would have to use a Bayesian deep learning modeling framework [401] where the constraints can be thought of as predicates that hold true in the data and correspondingly sampling from distributions using MCMC techniques. For the second choice, we would need to manually augment data by defining the way

counterfactuals are generated using labeling functions [351]. Comparing the evaluations of these two approaches are not trivial for domain experts without working with an extensive machine learning engineering team.

Our motivation is thus to enable a future of no-code or low-code abstractions for domain experts to incorporate their domain knowledge in the machine learning pipeline. To this end, we develop a specification language for domain experts to articulate how their domain knowledge can be mapped to several components like regualarization, data augmentation, model priors. To continue the running example of the medication recommendation task, using our framework, a doctor would be able to crisply define their domain knowledge in terms of the concepts they understand, and effectively map out which parts of the machine learning pipeline the said knowledge impact either in terms of loss functions, data distributions or model hypothesis space. For example, if the doctor had certain preferred medications for patients presenting with dermatological conditions on the hand, they could specify that using a mapping rule, and within the properties of that rule mention that this rule should hold over data augmentations performed on the fly using a custom loss function for concept-based regualarization loss. Each of the modules for data augmentation and regularization is readily available to the doctor with parameters they can choose to tweak as and when their data distributions and priors continue to evolve. Our framework then translates their specification into a control-flow graph of a standard machine learning framework like TensorFlow, by using just-in-time compilation optimizations to include the constraints specified, and approximating them where relevant (for e.g. non-differentiable functions, etc). Additionally, once these constraints are specified, our framework provides feedback to the domain expert as to the extent to which the constraints they specified are being satisfied during training and if any changes to the parameters are suggested based on sensitivity analysis. Thus, our framework provides domain experts visibility into the entire machine learning pipeline with the ability to intervene on each of the modules of the pipeline through programmatic specifications.

13.3 Related Work

Trustworthy ML: In classifiers, trust has been developed through enabling counterfactual explanations [140, 302, 327] and improving robustness in output predictions when inputs have imperceptible and label-invariant perturbations [199, 476, 212]. However, in real-world applications, making input changes that are imperceptible and label-invariant is difficult. While making models robust against these adversarial failure modes is important, they are orthogonal in scope for the use-cases domain experts care about. On the other hand, strictly enforcing fine-grained behaviors such as individual user-interaction safety guarantees [226], trust modeling [277] can be hard to achieve and further exacerbated by cold-start problems [33].

Hybrid Systems: Many approaches have been proposed to aid the domain expert in interpreting the machine learning model's predictions [136, 419]. Tools to guide the underlying deep learning model through interactive feedback [53] and inductive logic [439] that increases diversity and aligns the model's predictions to expert knowledge have been proposed in the medical domain [300]. Applying data mining to extract association rules using Bayesian methods between input and output categories are also well studied [240], but they are typically not validated with rules by experts.

Interpretability: Mapping human interpretable rules with ML models has also been done to understand the inner workings of a black box machine learning model. For a broad review of the various notions of interpretability, we refer to [97]. Our work closely relates to the "task related latent dimensions of interpretability". Here, we care about the hypothesis of local interpretability[355], with incomplete coverage of domain expertise [462]. By restricting to this type of interpretability over expert-defined rules on subsets of the data, we seek that our models obey those rules.

Adversarial Robustness: To make ML models robust to perturbations, prior work has proposed defenses so that the model does not change it's output prediction for a small (ϵ), but hu-

manly imperceptible change in the input [70, 54]. However, such adversarial robustness may either increase [188] or decrease [459] the overall accuracy of the models depending on the human specified notion of robustness. Hence, in the field of computer vision, robust models over concept based perturbations [445] and in natural language processing [185], robustness over word substitutions with synonyms are desired [344]. This indicates that the range of perturbations over which the robustness is defined, is equally important and going beyond geometrical definitions of robust boundaries is valuable [248, 346]. Hence, we choose to ground our models in expert defined relationships between inputs and outputs, which we would expect the non-observed data to generalize over.

Robustness in ML: Recently, there has been a lot of interest in making ML systems robust to avoid extremely undesired outcomes (e.g. horror films to children) [424, 448]. Robust ML models that explicitly guard against multiple attack models [187] like profile injection [82], noisy ratings [311] and implicit issues like outliers [397], data not missing at random [230] have been proposed. Our definition is complementary to prior work in robust recommender models which propose simpler models like decision trees [227], fairness guarantees to avoid unintended bias [35, 83, 384, 40], temporal coherence to avoid catastrophic forgetting [424], defense against adversarial attacks of imperceptible changes [66, 174], and uncertainty based model calibration [448]. However, such approaches implicitly assume the presence of embeddings of items on which a similarity function (e.g cosine similarity) can be applied and assign a penalty if the recommender predicts items with low similarity. Instead, we explicitly use domain specific rules defined over categories of items and expect that the recommendations do not deviate categorically from those rules. Additionally, such approaches focus primarily on training-time attacks and do not address counterfactual scenarios that might arise during inference.

Bayesian methods: Often domain knowledge imposes hard constraints on a model in the sense that the constraints exclude parts of the sample space altogether rather than only make them less likely. This means the target distribution may have isolated modes and sharp deriva-

tives, causing poor sampler exploration and instability in gradient-based methods. In these challenging cases, approximate inferences can be applied to *approximate models*. For instance, simulated annealing [224] can be used to bring the posterior closer to a uniform density, potentially connecting isolated modes. Parallel tempering samples from a collection of exact and approximate models (*replicas at increasing temperatures*) in parallel [395, 138, 102, 9, 396]. Predicate Exchange [401] applies parallel tempering to cases where sets are defined by hard predicates expressing arithmetic constraints, e.g. the square of one parameter is less than another parameter. To accomplish this, a set of atomic predicates (e.g. equality, inequality) are relaxed at each temperature to a function in $[0, 1]$, called a soft predicate, measuring the extent to which the constraint is satisfied. Complex soft constraints are then obtained recursively using relaxations of propositional connectives. However, the resulting specification language remains limited both in terms of expressivity and its one-choice-fits-all relaxation policy.

13.4 Domain Faithful Specifications

We build on this body of work to allow for abstract functional forms of constraints which allows domain experts to include free-form relationships between inputs and outputs. Specifically, we build on top of the just-in-time compilation and acceleration frameworks such as JAX [122] to encode robustness constraints as customized differentiable abstractions which can be added to traditional deep learning frameworks built in Tensorflow. This way, our expressivity is strictly better than relaxation based frameworks and can encode conditionals, loops, and other custom differentiable array operators. This choice of abstraction allows us to directly encode the constraints the domain experts wish to see in the model as additional augmentation and regularization losses, instead of a conditional sampling of model parameters. Further, by using decomposed reverse-mode automatic differentiation, we overcome the complexity of performing reverse-mode automatic differentiation of sampling algorithms like Hamiltonian Monte Carlo

and Metropolis Hastings over proposals on pre-defined proposals on elements in the dictionary. Instead of using fine-tuning temperature to swap between soft predicate proposals, we directly encode the underlying declarative knowledge in the form of constraints approximated by regularization penalties which allows to be backpropagated seamlessly. The difficulties brought on by the discontinuities in conditionals imposed by domain knowledge can be implemented directly using primitives such as "switch", "cond", "where", which allows for tracing of functions to track the shapes of the inputs and outputs, while still allowing for custom derivative rules over user-defined functions with discontinuities and non-differentiable points, leveraging the benefits of combining forward-mode residuals in reverse-mode automatic differentiation.

For example, given a constraint f, which is a functional mapping $f : C \rightarrow C$ between concepts $C = \{c_1, c_2, ..c_n\}$ that should be satisfied for any instance in the data x for which f is applicable. This can be represented as a lambda expression using a satisfiability function over the machine learning model M. If we assume that the membership of an instance x and it's output in the desired concepts is determined by functions i and o respectively, the domain set of concepts in f: D, and the correspondingly mapped concept set: v, then the functional specification can be written as follows:

Listing 13.1: Constraint functional specification

```
def satisfy(M, x):
    val = True
    if i(x) in D:
        val = o(M(x)) in v(x)
    return val
f = lambda x: satisfy(M, x)
```

Such a satisfiability constraint when specified can be incorporated into a framework like JAX as outlined in Section 13.5. One of the restrictions placed however is that the constraint needs to

be self-contained and not rely on any side-inputs which might change during execution. Hence, this supports a rich set of constraints beyond predicate first-order logic and enable higher-order logics over sets of data. This is specifically useful when constraints like statistical group fairness which operates over sets of data are required in a domain. However, in terms of expressability, we cannot specify a complete Turing machine as a constraint, due to our limitation on accessing side-inputs. However, the framework allows us to build new core primitives which allow calls into other functions, simulators, to be incorporated as long as the shapes and types of the inputs and outputs are statically determinable. By allowing each of these dynamic functions to be converted to core primitives and by allowing for abstract evaluation to determine the shapes and types of the arguments, we can efficiently run these functions over batches of data during training or "vectorize" them for just-in-time compilation and asynchronous dispatch. Conditionals whose shape and type are dynamic and based on the inputs are not amenable to this formulation and needs to be converted into statically analyzable primitives.

No-code specification library: To enable domain practitioners with little-to-no experience in coding, we also provide a JSON like specification interface for each of the above functions (see listing **??**). By tying this interface to the python specification and the primitive implementations of the constraints in the deep learning system, we provide an interpretable mechanism for the domain expert to evaluate the machine learning model with respect to their domain constraints.

13.5 System Design

In this section we describe the implementation of domain faithful deep learning system. Our implementation extends on JAX [122], and allows domain experts to leverage the benefits of just-in-time compilation which allows the execution of functional programs as part of the control flow graph. To allow for the full expressivity of domain declarative knowledge, we extend on the primitives built into JAX like tuples, dictionaries, along with nested data structures whose

gradients can be custom defined. Each deterministic constraint expressed functionally is then translated into a lambda JAX expression, which takes one or more input typed parameters with one or more typed results. Currently, JAX expressions do not support free variables within enclosing scopes. Further, the translation ignores any functional code, that is not translatable to a JAX expression. This allows that each of the modules can be written to capture domain specific knowledge in Python-level control flow, but only the relevant constraints that operate over inputs and outputs get translated into JAX Expressions. Each JAX Expression follows the below template:

Listing 13.2: JAX expression template

```
jaxpr ::= { lambda Var* ; Var+.
        let Eqn*
        in  [Expr+] }
```

Here, the first variables $Var*$ are dedicated for constant variables, $Var+$ denotes input variables, $Eqn*$ denotes a list of lambda expressions, which finally return the output variables.

13.5.1 Regularization

For example, the functional specification of a constraint that enforces that if one of the input is negative, then the output should be False which is expressed in a functional form, is then translated into a JAX expression (see Appendix, listing 13.6). Since the JIT compilations of JAX expressions are such that they can added onto a control graph of a traditional deep learning system and allows batching, we can then directly augment the constraint code to an existing loss function as follows:

Listing 13.3: Regularization Loss in DFS

```
def loss(W, b):
    preds = predict(W, b, inputs)
    satisfy = constraint(inputs, preds)
    label_probs = preds * targets + (1 - preds) * (1 - targets)
    return -jnp.sum(jnp.log(label_probs)) + jnp.sum(satisfy)
```

This allows for flexibility in how we define our constraints, the way in which the corresponding penalty terms are added to the base loss function (e.g log loss likelihood). Each of these choices have been parameterized for ease of specification by the domain expert.

13.5.2 Data Augmentation

Similar to regularization, data augmentation primitives can also be compiled JIT based on guidance from domain expert. For example, if augmentation transformations on structured data in the medication recommendation task is to replace one symptom with another symptom from the same category using our library "aug", we can simply write a batched augmentation pipeline to the existing batch of patient data as follows:

Listing 13.4: Data Augmentation in DFS

```
import jax
import dfs

transform = dfs.Chain(
  dfs.RandomCategoryChoice(),
  dfs.SwapInput(),
  dfs.PopulateCategoryLabel(),
)
```

```
rng = jax.random.PRNGKey(42)
sub_rngs = jax.random.split(rng, patients.shape[0])
aug_patients = jax.jit(jax.vmap(transform))(sub_rngs, patients)
```

By doing this transformation, we can generate augmented data just-in-time, and also incorporate expert domain knowledge in doing so. For example, if experts want to peform data augmentation which allows only certain types of input swaps, we encode that logic in the function "SwapInput". Here, too we provide a library of commonly data augmentation techniques for the domain expert to choose between.

13.6 Properties of Domain Faithful Deep Learning

13.6.1 Evaluating Safety

Each of the JAX expressions can be custom evaluated as part of the compilation to allow for domain-specific interpretation of the model's performance. For example, in order to evaluate how often the above example constraint is violated, we write a custom interpreter for the above JAX expression, which evaluates and reports the number of violations of the constraint using an auxiliary reporting output variable. This wrapping mechanism has further been automated using custom decorators called "@domaineval", which correspondingly reports such behaviors using wrapper libraries like "flax". Here too, we have built a library of relevant mappings and the corresponding metrics that would be useful for the domain expert based on prior user studies.

```
@cons(name='negativeInput')
class StrictModule(nn.Module)
    @domaineval(allowedRate)
    def __call__(inputs, preds):
```

```
        satisfy = constraint(inputs, preds)
        self.sow('intermediates', 'constraintRate', satisfy)
        return jax.nn.celu(satisfy, 1/allowedRate)
```

13.6.2 Parameter Optimization

The specification of the domain specification library encapsulates all parameters relevant to a constraint in a dictionary. This allows the domain experts to train domain faithful end-to-end neural networks by simply specifying the parameters of their constraints.

Listing 13.5: Functional specification for category concordance

```
exp = {
constraints = [
    constraint {
        name = 'negativeInput'
        allowedRate = 0.1
        onFailure = None
        evalReport = True
    }
],
preprocessing = [
    augment {
        name = 'categorySwap'
        trainingRatio = 0.2
    }
],
```

```
input = './path/to/input',
output = './path/to/output',
classification = 'crossEntropy',
loss = 'logLoss',
epochs = 10,
}
```

Since improving model robustness as defined by the domain constraints is the primary objective of DFS, we assume that the values of the constraint parameters need to be continuously fine-tuned. In DFS, we have also implemented a mechanism to provide feedback on the numerical values, using techniques of Bayesian optimization [274] which models the prior of the values based on the domain expert assigned values and evaluates based on an additive semiparametric error $\epsilon(x)$ to compute the posterior of the surrogate model $P(f|D)$ given an initial distribution of functions $P(f)$ (a Gaussian process) and a criterion to acquire new values. In our empirical Bayes implementation, we use the acqusition function based on a combination of global and local derivate-free method - BOBYQA [336]. This feature can be simply enabled by mentioning two parameters in the above specification - the start and stop epoch number when the constraint parameters will be optimized.

13.7 Evaluation

13.7.1 Discontinuous Constraints

We use the histograms of samples from a uniform prior $[-1, 1]^2$ used in [401], conditioned on a variety of predicates. These examples are simple yet challenging to simple feedforward neural networks to learn due to discontinuities in the approximate posterior. Specifically, the true distribution is from $x, y \sim Unif(-1, 1)$ with conditions of $x = y$, $|x| \geq |y|$, $x^2 = y^2$ and

Model Version	RMSE to ground truth
Baseline	0.84 (0.80, 0.88)
Baseline+Mapped	0.75 (0.74, 0.76)
Baseline+Predicate	0.72 (0.68, 0.76)
RA	0.65 (0.63, 0.68)
RA-WCR	0.42 (0.41, 0.43)
RA-WCR [dfdl]	**0.28 (0.21, 0.35)**

Table 13.1: Our method considerably reduces the RMSE to the ground truth on average over the 4 synthetic conditioned models.

$sin(kx).cos(kx) \geq 0.9999$. In each of these cases, we are given a partial domain knowledge in the space of $x, y \sim Unif(-0.1, 0.1)$ where each of the conditions and the corresponding constraints are known to be true. By adding an augmentation module to each of these distributions, our models are expected to generalize beyond the space where the constraints are known to be true. Further, the continuous refinement of the thresholds of the constraints, further indicate the growing boundaries within which the constraints hold true. This further indicates to the domain expert to incrementally increase the boundaries and finally end up in a much more generalized model as shown in Table 13.1. Baseline is a feedforward network model with two layers with 10 hidden units with non-linear units (RELU) run for 100 epochs.

13.7.2 Medication Recommendation Case Study

With the advent of Electronic Health Record(EHR), doctors are able to make significantly better clinical decision with the help of rule based recommendation systems.

Given a set of mappings between categories of diagnoses C_X and categories of medication C_Y, we want to learn a neural network model f which consistently maps diagnoses $x : c_x \in C_X$ to their corresponding medications $y : c_y \in C_Y$ as per the mapping, such that $\hat{y} = f(x)$. Let us define a function p which maps the diagnoses categories to medication categories i.e $p : C_X \rightarrow C_Y$ $\forall x : c_x \in C_X, c_{\hat{y}} = p(c_x)$.

We follow the MIMIC-III medication recommendation task as per [376], and the domain specific mappings p are obtained from [**atc-icd**] where medical experts validated a statistical table

based on pairwise mutual information scores of co-occurrences between diagnostic x (ICD-9) and medication y (ATC) codes. These validated tables are segmented based on the age and gender of Austrian patients. Note that this dataset is different from the MIMIC-III dataset used in our evaluation. Hence, we use only the pairs of ICD-9: j, ATC categories: k that are expert validated p, but not any other statistical information from this study. A total of unique 349 pairs of ATC and ICD-9 Level 2 codes were deemed to be valid by the experts; 958 unique pairs if we break down by age and gender forms our domain specific mapping p. Age is bracketed into 3 ranges based on year of birth (1949-68, 1969-88, 1989-2008) and gender is considered to be binary (male, female). The categorical distance d_c used to define the robustness distance is given by the path distance between ICD-9 codes in the ICD-9 ontology tree. We use these validated pairs to generate perturbations in our existing dataset. The constraints used are category misclassification loss $\mathcal{L}_v(D_p)$ and the within-category regularization loss $\mathcal{L}_{ar}(D_p)$ in this domain:

$$\mathcal{L}_v(D_p) = \mathbb{E}_{(X,Y)\in D_p} \mathop{\mathbb{E}}_{\substack{(j,k):p(j)=k \\ X'\sim\delta_j(X)}} \mathbb{I}(k \notin \bigcup_{y'\in h(X')} f_O(y')) \tag{13.1}$$

$$\mathcal{L}_{ar}(D_p) = \mathop{\mathbb{E}}_{\substack{(X,Y)\in D_p,(j,k):p(j)=k \\ X'\sim\delta_j(X),y\sim Y\cap\mathcal{Y}_k}} \mathcal{L}_r(X, X', y) \tag{13.2}$$

13.7.3 Baseline Models

We use the current state-of-the-art for the medication recommendation task on MIMIC-III dataset as the *Baseline* - G-BERT [376]. This model uses graph embeddings based on the ontology of the ATC and ICD-9 codes. The model initially pre-trains the embeddings on the single-visit data using self-supervised learning, similar to BERT [85]. The graph embeddings are learnt using the

Graph Attention technique [416], so as to learn hierarchical embeddings for each of the diagnostic and medication codes.

For each of these domains, we define the current state-of-the-art model as ***Baseline***. As our framework incorporates more information through robust domain specific mappings through counterfactual augmented data, we also developed additional baselines that used these priors as input features. Specifically, we augmented categorical embeddings of each input to form the ***Baseline*+Cat** model. In this baseline, no expert validation information is provided, but the category embedding is explicitly provided. We also augmented the embeddings of the applicable rule-based output category $k : p(j) = k$ as an input to the model to form the ***Baseline*+Mapped** model. This trains the model to pay attention to the mapped output category and minimize category misclassification. We also incorporate the predicate exchange algorithm [401] to incorporate the categorical rules as stochastic approximations which is then fine-tuned using a temperature-scaled MCMC to draw posterior samples for a probabilistic program execution. Finally, we instantiate our models ***Baseline* RA**, which modifies the baseline with Rule-based Augmentation (Eq. 13.1) and ***Baseline* RA-WCR**, which uses Rule-based Augmentation and Within-Category Regularization (Eq. 13.2). Our domain faithful deep learning version implemented in JAX is called We set $\alpha = 0.2$ after cross-validation.

13.7.4 Metrics

To build the neural models that follow domain rules, we regularize the model such that within-category loss (13.2) is minimized. We evaluate improvement in robustness using the following distance metric between inputs.

Definition 13.1. Robustness Distance: Given all rules of the form $p(j) = k$, and the subset of the dataset D covered by them: D_p, *robustness distance* is measured as the average of the minimum categorical distance d_c between input categories j and j', where $x : j \in f_l(x)$ and a single item

perturbation $x' \in S_k(X) : j' \in f_I(x')$ that leads to k being removed from the set of perturbed output categories $O(X')$.

$$O(X') = \{f_O(y') : \forall y' \in h(X')\} \tag{13.3}$$

$$S_k(X) = \{x' | X' = x' \cup X \setminus x \wedge x \in X \wedge k \notin O(X')\} \tag{13.4}$$

$$d_{robust} = \mathbb{E}_{(X,Y)\in D_p}[\min_{\substack{j\in f_I(x), j'\in f_I(x') \\ x\in X, p(j)=k, x'\in S_k(X)}} (d_c(j, j'))] \tag{13.5}$$

13.8 Results

13.8.1 Accuracy

To test if we improve accuracy on the original dataset, we evaluate overall accuracy metrics. For the medication recommendation task as shown in Table 13.2, in the MIMIC-III diagnostic code classification task through domain faithful deep learning parameter refinements, we *improve F1-score by 5.3%* with similar gains in Jaccard coefficient and PR-AUC and we *improve F1-score by 3.3%* on the medicine category classification task over the augmented dataset which contains counterfactual scenarios of in-category diagnostic codes, thereby increasing adherence to diagnostic-medication category mappings.

13.8.2 Domain Robustness

We now test: "Does our method effectively increase adherence to the domain experts' mappings?" To measure if neural recommender models follow domain-specific rules, we evaluate the robustness distance as defined in *Definition 13.1*, limited to the subset of the data specified by the mappings. To continue the ICD-9 code based medication recommendation example, the

	Model	Jaccard	F1	PR-AUC
Original	G-Bert	0.3679 ±0.01	0.5281 ±0.03	0.6212 ±0.03
	G-Bert+Cat	0.3564 ±0.02	0.5203 ±0.04	0.6146 ±0.03
	G-Bert+Mapped	0.3680 ±0.01	0.5299 ±0.03	0.6230 ±0.02
	G-Bert+Predicate	0.3470 ±0.01	0.5149 ±0.03	0.6291 ±0.02
	G-Bert RA	0.3883 ±0.02	0.5788 ±0.02	0.6541 ±0.01
	G-Bert RA-WCR	0.4300 ±0.01	0.5967 ±0.01	0.6775 ±0.02
	G-Bert RA-WCR [dfdl]	**0.4530** ±0.01	**0.6132** ±0.01	**0.6802** ±0.02
Augmented	G-Bert	0.3677 ±0.03	0.5281 ±0.02	0.6199 ±0.00
	G-Bert+Cat	0.3301 ±0.03	0.5102 ±0.01	0.5952 ±0.01
	G-Bert+Mapped	0.3573 ±0.01	0.5249 ±0.02	0.6084 ±0.02
	G-Bert+Predicate	0.3564 ±0.01	0.5223 ±0.02	0.6051 ±0.02
	G-Bert RA	0.3723 ±0.02	0.5483 ±0.02	0.6343 ±0.01
	G-Bert RA-WCR	0.4033 ±0.01	0.5699 ±0.02	0.6596 ±0.02
	G-Bert RA-WCR [dfdl]	**0.4127** ±0.01	**0.5742** ±0.02	**0.6621** ±0.02

Table 13.2: Our RA-WCR model improves accuracy metrics of G-BERT on the MIMIC-III medication recommendation task after fine-tuning the parameters of the constraints for the Original dataset and the category classification task for the within-category Augmented dataset

changes would be quantified by the edge distance in the ICD-9 code ontology required to change the output ATC medication code. As shown in Table 13.3, our G-BERT RA-WCR model *achieves a robustness distance* $d_{robust} = 2.8 \geqslant 2.4 \geqslant 2$, suggesting that the model on average follows the expert-defined rules for counterfactuals near observed examples as compared to doing the augmentation and robustness regularization manually. Having a robustness distance greater than or equal to 2, *implies* that on average for any change in the recommended medication category, the model expects that the input diagnostic code category should have also changed.

13.9 Conclusion

In this paper, we have outlined a domain faithful deep learning framework based on the just-in-time compilation and abstract evaluation and tracing of the control flow of deep learning systems. We leverage the benefits of this decoupling between functional specifications and batched exe-

Model Version	Baseline: G-BERT (MIMIC-III) d_{robust} (ICD-9 tree distance)
Baseline	1.3 (1.0, 1.6)
Baseline+Cat	1.1 (1.0, 1.2)
Baseline+Mapped	1.2 (1.0, 1.4)
Baseline+Predicate	1.2 (1.0, 1.4)
RA	1.7 (1.5, 1.9)
RA-WCR	2.4 (2.1, 2.7)
RA-WCR [dfdl]	**2.8 (2.7, 2.9)**

Table 13.3: Our method considerably increases the mean robustness distance (± standard deviation in brackets - see Def. 13.1) in the medication domain.

cution afforded by our framework, to enable continuous refinement of parameters by domain experts, including hints to change hyperparameters of the constrains and augmentation modules to optimize for overall accuracy and robustness in both synthetic and MIMIC-III medication recommendation tasks.

13.10 Appendix

```
import jax.numpy as jnp
def constraint(inputs, preds):
    return jnp.where(
        jnp.take(inputs, 2, 1) < 0,
        preds, jnp.array([0]*len(preds)))
```

Listing 13.6: JAX Expression for Functional Specification

```
invars: [b, c]
outvars: [f]
constvars: [a]
equation: [b, 2] xla_call [d] {'device': None, 'backend': None, 'name':
```

```
    c:bool[] = lt b 0
    d:i32[] = add b 3
    e:i32[] = xla_call[
      call_jaxpr={ lambda ; f:bool[] g:i32[] h:i32[]. let
          i:i32[] = select_n f h g
        in (i,) }
      name=_where
    ] c d b
    j:i32[1] = broadcast_in_dim[broadcast_dimensions=() shape=(1,)] e
    k:f32[4] = gather[
      dimension_numbers=GatherDimensionNumbers(offset_dims=(0,), collaps
      fill_value=None
      indices_are_sorted=False
      mode=GatherScatterMode.CLIP
      slice_sizes=(4, 1)
      unique_indices=False
    ] a j
  in (k,) }}
equation: [d, 0.0] lt [e] {}
equation: [e, c, a] xla_call [f] {'device': None, 'backend': None, 'nam
    d:i32[4] = convert_element_type[new_dtype=int32 weak_type=False] b
    e:i32[4] = select_n a c d
  in (e,) }}
```

14 | Conclusion

Through this research, we have established several foundational implications for various application domains. Through incorporating causal models in natural language processing models, we have seen that causal graph faithfulness can lead to better generalization across domains, incorporating indicative causal features in time series predictive models can lead to better modeling of heterogeneous effects in spatio-temporal structured prediction social science tasks, generating data that models nuanced counterfactual behavior can provide robustness guarantees when evaluated against medical ontological mappings, identifying when covariate overlap assumptions are invalid can better eliminate spurious temporal correlations, interpreting and improving sliced accuracy across demographic groups can improve the pareto frontier of machine learning models, and finally it is possible for domain experts to specify their domain knowledge in a manner which is both easy to represent but also optimized for training large scale deep learning models. Through each of these insights, our work also opens up a wide area of research to explore in the future.

Specifically, developing a framework where domain experts and ML practitioners can collaborate on mutually beneficial abstractions for fairness, concordance, causal models, etc. that is interpretable for the practitioners and operable for the ML researchers is an area that can be extensively studied. Such Domain Faithful Deep Learning systems will be flexible to various types of domain knowledge including but not limited to categorical mappings, logical formulations over concepts, algebraic constraints over groups of data. With this framework, we have explored

building Domain Faithful Deep Learning Systems with applications in the realm of healthcare, sustainability, responsible computational social science and privacy by addressing the following core challenges.

- **Domain specification language**: One of the hurdles to enable such systems is the lack of a common specification language for practitioners and researchers to collaborate. For example, in the medication recommendation task, have automated the process of data augmentation, regularization, into a specification language for medical domain experts. This not only improves the transparency of ML design, but allows researchers more flexibility in choosing among techniques applicable for the health domain.

- **Domain structure for global properties**: Incorporating global properties over large groups of data instances into ML models needs to be an integral part of design choices in trustworthy socio-technical systems. Our work on improving pareto efficiency and sliced accuracy through secondary proxy variables has shown that we can build robust models for all demographic groups while improving overall accuracy.

- **Scientific Hypotheses Discovery**: Further, in many domains where domain knowledge is still in its nascent phase, our research has been used to analyze the performance of the ML models while keeping domain specific constraints in mind, which can pave the way for generating hypotheses for scientific discovery. Causal model discovery and extraction as evidenced in the famine forecasting work, has shown that accurate and robust models can be built with existing scientific hypotheses in mind.

- **Translating natural language for logical applications**: Similarly, domains which have complex unstructured data can benefit from using ML to incorporate domain structure to be checked by experts. For example, in the domain of question answering, we have shown that by using causally faithful representations, we can directly deploy logically motivated graphs in information retrieval systems.

- **Ethical translation of domain knowledge**: How domain expertise gets translated into statistical constraints and concepts can have ethical implications. The questions such as what data distribution and for what purpose is the model trained intended for, are closely related and is precisely the type of cross-disciplinary analysis we need to engage domain experts with, for building responsible data-driven systems. For example, in the coronary angiographic disease status prediction task and outcome-aware agricultural household clustering, how we balance the error rates across demographic groups can unearth historical biases in the measurements and calibrations of the data. This way, we have incorporated socio-economic inference models as part of participatory policy making and algorithmic decision making.

Through our research, we have enabled domain experts and ML researchers to work together and converge to a common understanding of how the ML models operate. The challenges of the future like climate change, pollution, health and toxicity in social media need our concerted efforts. Through this research on incorporating domain structure into end-to-end ML models, we have opened the doors for domain experts like economists, doctors, physicists, gene biologists, earth scientists, linguists, lawyers and social scientists to provide inputs based on their domain knowledge to help build robust ML models in their fields for safe decision making.

Bibliography

[1] Nabeel Abdur Rehman et al. "Fine-grained dengue forecasting using telephone triage services". In: *Science Advances* 2.7 (2016). DOI: 10.1126/sciadv.1501215. eprint: http://advances.sciencemag.org/content/2/7/e1501215.full.pdf. URL: http://advances.sciencemag.org/content/2/7/e1501215.

[2] Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. "Learning Edge Representations via Low-Rank Asymmetric Projections". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Nov. 2017). DOI: 10.1145/3132847.3132959. URL: http://dx.doi.org/10.1145/3132847.3132959.

[3] A Agarwal et al. "A Reductions Approach to Fair Classification". In: *CoRR* abs/1803.02453 (2018). arXiv: 1803.02453. URL: http://arxiv.org/abs/1803.02453.

[4] David Ahn. "The stages of event extraction". In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 1–8. URL: https://www.aclweb.org/anthology/W06-0901.

[5] Betty van Aken et al. *Challenges for Toxic Comment Classification: An In-Depth Error Analysis*. 2018. arXiv: 1809.07572 [cs.CL].

[6] J. C. Aker. "Comparing cash and voucher transfers in a humanitarian context: Evidence from the Democratic Republic of Congo". In: *The World Bank Economic Review* 31.1 (2017), pp. 44–70.

[7] J Ali et al. “Loss-Aversively Fair Classification”. In: AIES ’19. Honolulu, HI, USA, 2019, pp. 211–218. ISBN: 9781450363242. DOI: 10.1145/3306618.3314266. URL: https://doi.org/10.1145/3306618.3314266.

[8] A. Allabban, JE Hollander, and JM Pines. “Gender, race and the presentation of acute coronary syndrome and serious cardiopulmonary diagnoses in ED patients with chest pain”. In: *Emergency Medicine Journal*. Vol. 34. 2017, pp. 653–658.

[9] Gautam Altekar et al. “Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference”. In: *Bioinformatics* 20.3 (2004), pp. 407–415.

[10] Moustafa Alzantot et al. “Generating Natural Language Adversarial Examples”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2890–2896. DOI: 10.18653/v1/D18-1316. URL: https://www.aclweb.org/anthology/D18-1316.

[11] U. Amarasinghe, M. Samad, and M. Anputhas. “Spatial clustering of rural poverty and food insecurity in Sri Lanka”. In: *Food Policy* 30.5-6 (2005), pp. 493–509.

[12] Giuseppe Amodeo, Roi Blanco, and Ulf Brefeld. “Hybrid Models for Future Event Prediction”. In: CIKM ’11. 2011, pp. 1981–1984.

[13] Bo Pieter Johannes Andrée. “Estimating Food Price Inflation from Partial Surveys”. In: *World Bank Policy Research Working Papers* (2021).

[14] Bo Pieter Johannes Andrée et al. “Predicting Food Crises”. In: *World Bank Policy Research Working Papers* (2020).

[15] Martin Arjovsky et al. *Invariant Risk Minimization*. 2020. arXiv: 1907.02893 [stat.ML].

[16] Andrew Arnold, Yan Liu, and Naoki Abe. “Temporal Causal Modeling with Graphical Granger Methods”. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’07. San Jose, California, USA: ACM, 2007,

pp. 66–75. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281203. URL: http://doi.acm.org/10.1145/1281192.1281203.

[17] Andrew Arnold, Yan Liu, and Naoki Abe. "Temporal Causal Modeling with Graphical Granger Methods". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 66–75. ISBN: 9781595936097. DOI: 10.1145/1281192.1281203. URL: https://doi.org/10.1145/1281192.1281203.

[18] L. Arockiam, S. S. Baskar, and L. Jeyasimman. "Overview of clustering techniques in agriculture data mining". In: *Agricultural Journal* 6.5 (2011), pp. 222–225.

[19] Nabiha Asghar. "Yelp Dataset Challenge: Review Rating Prediction". In: *CoRR* abs/1605.05362 (2016). arXiv: 1605.05362. URL: http://arxiv.org/abs/1605.05362.

[20] Henry Assael. "Segmenting markets by group purchasing behavior: an application of the aid technique". In: *Journal of Marketing Research* (1970), pp. 153–158.

[21] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects: Table 1." In: vol. 113. July 2016, pp. 7353–7360. DOI: 10.1073/pnas.1510489113.

[22] S. Auer et al. "DBpedia: A Nucleus for a Web of Open Data". In: *ISWC'07/ASWC'07*. Busan, Korea, 2007, pp. 722–735.

[23] David Backer and Trey Billing. "Validating Famine Early Warning Systems Network projections of food security in Africa, 2009–2020". In: *Global Food Security* 29 (2021), p. 100510. ISSN: 2211-9124. DOI: https://doi.org/10.1016/j.gfs.2021.100510. URL: https://www.sciencedirect.com/science/article/pii/S2211912421000201.

[24] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. "The Berkeley FrameNet Project". In: (Aug. 1998), pp. 86–90. DOI: 10.3115/980845.980860. URL: https://www.aclweb.org/anthology/P98-1013.

[25] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. “The Berkeley FrameNet Project”. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. ACL ’98/COLING ’98. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998, pp. 86–90. DOI: 10.3115/980845.980860. URL: https://doi.org/10.3115/980845.980860.

[26] Ananth Balashankar et al. “Identifying Predictive Causal Factors from News Streams”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019. DOI: 10.18653/v1/D19-1238. URL: https://www.aclweb.org/anthology/D19-1238.

[27] A. V. Banerjee, E. Duflo, and M. Kremer. “The influence of randomized controlled trials on development economics research and on development policy”. In: *The state of Economics, the state of the world* (2016), pp. 482–488.

[28] Elias Bareinboim and Judea Pearl. “Causal inference and the data-fusion problem”. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7345–7352. ISSN: 0027-8424. DOI: 10.1073/pnas.1510507113. eprint: https://www.pnas.org/content/113/27/7345.full.pdf. URL: https://www.pnas.org/content/113/27/7345.

[29] Valerio Basile et al. “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://www.aclweb.org/anthology/S19-2007.

[30] D. A. Belsley, E. Kuh, and R. E Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. 1980.

[31] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://www.aclweb.org/anthology/2020.acl-main.463.

[32] Shane Bergsma and Dekang Lin. "Bootstrapping Path-Based Pronoun Resolution". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. ACL-44. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 33–40. DOI: 10.3115/1220175.1220180. URL: https://doi.org/10.3115/1220175.1220180.

[33] L. Bernardi et al. "The continuous cold start problem in e-commerce recommender systems". In: *arXiv:1508.01177* (2015).

[34] Ivo Bernardo, Roberto Henriques, and Victor Lobo. "Social Market: Stock Market and Twitter Correlation". In: *Intelligent Decision Technologies 2017*. Ed. by Ireneusz Czarnowski, Robert J. Howlett, and Lakhmi C. Jain. Cham: Springer International Publishing, 2018, pp. 341–356. ISBN: 978-3-319-59424-8.

[35] A. Beutel et al. "Fairness in Recommendation Ranking through Pairwise Comparisons". In: *KDD* (2019), pp. 2212–2220.

[36] Alex Beutel et al. "Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements". In: *CoRR* abs/1901.04562 (2019). arXiv: 1901.04562. URL: http://arxiv.org/abs/1901.04562.

[37] A Beutel et al. "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations". In: *CoRR* abs/1707.00075 (2017).

[38] Sangeeta Bhatia et al. "Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread". In: *NPJ digital medicine* 4.1 (2021), pp. 1–10.

[39] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley". In: *Science* 187.4175 (1975), pp. 398–404. ISSN: 0036-8075. DOI: 10.1126/science.187.4175.398. eprint: http://science.sciencemag.org/content/187/4175/398.full.pdf. URL: http://science.sciencemag.org/content/187/4175/398.

[40] A. J. Biega, K. P. Gummadi, and G. Weikum. "Equity of Attention: Amortizing Individual Fairness in Rankings". In: *SIGIR*. ACM, 2018, pp. 405–414.

[41] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022.

[42] Su Lin Blodgett et al. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 5454–5476.

[43] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata". In: *Science* 350.6264 (2015), pp. 1073–1076. ISSN: 0036-8075. DOI: 10.1126/science.aac4420. URL: https://science.sciencemag.org/content/350/6264/1073.

[44] Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. "Robustness to Capitalization Errors in Named Entity Recognition". In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 237–242. DOI: 10.18653/v1/D19-5531. URL: https://www.aclweb.org/anthology/D19-5531.

[45] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *CoRR* abs/1607.06520 (2016). arXiv: 1607.06520. URL: http://arxiv.org/abs/1607.06520.

[46] T Bolukbasi et al. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”. In: *CoRR* abs/1607.06520 (2016). URL: http://arxiv.org/abs/1607.06520.

[47] Luca Bombelli, Johan Noldus, and Julio Tafoya. *Lorentzian Manifolds and Causal Sets as Partially Ordered Measure Spaces*. 2013. arXiv: 1212.0601 [gr-qc].

[48] Stephen Bonner and Flavian Vasile. “Causal Embeddings for Recommendation”. In: *CoRR* abs/1706.07639 (2017). arXiv: 1706.07639. URL: http://arxiv.org/abs/1706.07639.

[49] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. 2nd. Springer, Mar. 2002. ISBN: 0387953515.

[50] Morton B Brown and Alan B Forsythe. “Robust tests for the equality of variances”. In: *Journal of the American Statistical Association* 69.346 (1974), pp. 364–367.

[51] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[52] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *FAT*. 2018.

[53] C. J. Cai et al. “Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making”. In: *CHI*. 2019.

[54] Y. Carmon et al. “Unlabeled data improves adversarial robustness”. In: *NeurIPS*. 2019, pp. 11192–11203.

[55] M. Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: *Ferrari V*. Ed. by M. Hebert, C. Sminchisescu, and Y. Weiss. vol 11218. Springer, Cham: Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, 2018.

[56] A Chakraborty et al. “Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations”. In: *CoRR* abs/1704.00139 (2017). arXiv: 1704.00139. URL: http://arxiv.org/abs/1704.00139.

[57] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. “Multi-level cause-effect systems”. In: *Artificial Intelligence and Statistics*. 2016, pp. 361–369.

[58] Ching-Yun Chang et al. “Measuring the Information Content of Financial News”. In: *COLING*. ACL, 2016, pp. 3216–3225.

[59] Irene Chen, Fredrik D. Johansson, and David A Sontag. “Why Is My Classifier Discriminatory?” In: *CoRR* abs/1805.12002 (2018).

[60] Mo Chen, Qiong Yang, Xiaoou Tang, et al. “Directed Graph Embedding.” In: *IJCAI*. 2007, pp. 2707–2712.

[61] Zhitang Chen et al. “Causal discovery via reproducing kernel hilbert space embeddings”. In: *Neural computation* 26.7 (2014), pp. 1484–1517.

[62] Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu. “FBLG: A Simple and Effective Approach for Temporal Dependence Discovery from Time Series Data”. In: KDD ’14. 2014, pp. 382–391. DOI: 10.1145/2623330.2623709.

[63] M. D. Choudhury and E. Kiciman. “The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk”. In: *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media* 2017 (2017), pp. 32–41.

[64] Alexandra Chouldechova and Max G’Sell. “Fairer and more accurate, but for whom?” In: (June 2017).

[65] Gerardo Chowell et al. “Synthesizing data and models for the spread of MERS-CoV, 2013: Key role of index cases and hospital transmission”. In: *Epidemics* 9 (2014), pp. 40–51. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2014.09.011. URL: http://www.sciencedirect.com/science/article/pii/S1755436514000607.

[66] K. Christakopoulou and A. Banerjee. “Adversarial Attacks on an Oblivious Recommender”. In: *RecSys*. Copenhagen, Denmark, 2019, pp. 322–330.

[67] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. “Don’t Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4069–4082. DOI: 10.18653/v1/D19-1418. URL: https://aclanthology.org/D19-1418.

[68] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. “Learning to Model and Ignore Dataset Bias with Mixed Capacity Ensembles”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3031–3045. DOI: 10.18653/v1/2020.findings-emnlp.272. URL: https://aclanthology.org/2020.findings-emnlp.272.

[69] CMU. *Carnegie Mellon University Motion Capture Database*. 2009. URL: http://mocap.cs.cmu.edu/.

[70] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. “Certified adversarial robustness via randomized smoothing”. In: *arXiv preprint arXiv:1902.02918* (2019).

[71] Robert F. Cohen et al. “Three-dimensional graph drawing”. In: *Graph Drawing*. Ed. by Roberto Tamassia and Ioannis G. Tollis. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 1–11. ISBN: 978-3-540-49155-2.

[72] Nigel Collier et al. “BioCaster: detecting public health rumors with a Web-based text mining system”. In: *Bioinformatics* 24.24 (2008), pp. 2940–2941. DOI: 10.1093/bioinformatics/btn534. URL: http://dx.doi.org/10.1093/bioinformatics/btn534.

[73] William Conover and Ronald Iman. “[Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics]: Rejoinder”. In: *American Statistician - AMER STATIST* 35 (Aug. 1981), pp. 124–129. DOI: 10.1080/00031305.1981.10479327.

[74] Sam Crowe et al. “A plan for community event-based surveillance to reduce Ebola transmission - Sierra Leone, 2014-2015”. In: *MMWR. Morbidity and mortality weekly report* 64.3 (Jan. 2015), pp. 70–73. ISSN: 0149-2195. URL: http://europepmc.org/articles/PMC4584562.

[75] Alexander D’Amour et al. *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. 2020.

[76] Wayne W. Daniel. “Kolmogorov-Smirnov one-sample test”. In: *Applied Nonparametric Statistics (2nd ed.)* (1990), pp. 319–330.

[77] Ali F. Darrat, Maosen Zhong, and Louis T.W. Cheng. “Intraday volume and volatility relations with and without public news”. In: *Journal of Banking and Finance* 31.9 (2007), pp. 2711–2729. ISSN: 0378-4266. DOI: https://doi.org/10.1016/j.jbankfin.2006.11.019. URL: http://www.sciencedirect.com/science/article/pii/S0378426607000568.

[78] Dipanjan Das et al. “Probabilistic Frame-Semantic Parsing”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, June 2010, pp. 948–956. URL: https://www.aclweb.org/anthology/N10-1138.

[79] Sumanth Dathathri et al. “Plug and Play Language Models: A Simple Approach to Controlled Text Generation”. In: *CoRR* abs/1912.02164 (2019). arXiv: 1912.02164. URL: http://arxiv.org/abs/1912.02164.

[80] Ashlynn R. Daughton et al. “An approach to and web-based tool for infectious disease outbreak intervention analysis”. In: *Nature Scientific Reports, volume 7, Article number: 46076* (2017).

[81] Gerard Debreu. “Valuation equilibrium and Pareto optimum”. In: *Proceedings of the National Academy of Sciences of the United States of America* 40.7 (1954), p. 588.

[82] Y. Deldjoo et al. “How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models”. In: *SIGIR*. Virtual Event, China, 2020, pp. 951–960.

[83] Y. Deldjoo et al. *Recommender Systems Fairness Evaluation via Generalized Cross Entropy*. 2019. arXiv: 1908.06708.

[84] Benedict Dempsey and Debbie Hillier. “A Dangerous delay: The cost of late response to early warnings in the 2011 drought in the Horn of Africa”. In: *Oxfam* (2012).

[85] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*. 2019.

[86] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[87] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[88] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[89] M. Dimakopoulou et al. "Estimation considerations in contextual bandits. arXiv". preprint. 2017.

[90] Xiao Ding et al. "Deep Learning for Event-Driven Stock Prediction". In: *IJCAI*. AAAI Press, 2015, pp. 2327–2333.

[91] Xiao Ding et al. "Knowledge-Driven Event Embedding for Stock Prediction". In: *COLING*. ACL, 2016, pp. 2133–2142.

[92] Xiao Ding et al. *Using Structured Events to Predict Stock Price Movement: An Empirical Investigation*. 2014.

[93] Y. Ding, J. Liu, and D. Wang. "Deep Feature Fusion over Multi-Field Categorical Data for Rating Prediction". In: *AICCC*. Tokyo, Japan, 2018, pp. 16–22.

[94] Lucas Dixon et al. "Measuring and Mitigating Unintended Bias in Text Classification". In: 2018.

[95] Lucas Dixon et al. "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.

[96] Quang Xuan Do, Yee Seng Chan, and Dan Roth. "Minimally Supervised Event Causality Identification". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 294–303. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145466.

[97] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[98] P. van den Driessche and James Watmough. “Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission”. In: *Mathematical Biosciences* 180.1 (2002), pp. 29–48. ISSN: 0025-5564. DOI: https://doi.org/10.1016/S0025-5564(02)00108-6. URL: http://www.sciencedirect.com/science/article/pii/S0025556402001086.

[99] Cynthia Dwork and Christina Ilvento. “Fairness Under Composition”. In: *ITCS*. 2019.

[100] Cynthia Dwork et al. “Fairness through awareness”. In: *ITCS*. 2012.

[101] C Dwork et al. “Decoupled Classifiers for Group-Fair and Efficient Machine Learning”. In: ACM FAccT. 2018, pp. 119–133.

[102] David J Earl and Michael W Deem. “Parallel tempering: Theory, applications, and new perspectives”. In: *Physical Chemistry Chemical Physics* 7.23 (2005), pp. 3910–3916.

[103] E. E. Eban et al. “Scalable Learning of Non-Decomposable Objectives”. In: *ArXiv e-prints* (Aug. 2016). arXiv: 1608.04802 [stat.ML].

[104] Javid Ebrahimi et al. “HotFlip: White-Box Adversarial Examples for Text Classification”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 31–36. DOI: 10.18653/v1/P18-2006. URL: https://www.aclweb.org/anthology/P18-2006.

[105] Danielle Ensign et al. “Runaway Feedback Loops in Predictive Policing”. In: *CoRR* abs/1706.09847 (2017). arXiv: 1706.09847. URL: http://arxiv.org/abs/1706.09847.

[106] Alessandro Epasto and Bryan Perozzi. “Is a Single Embedding Enough? Learning Node Representations that Capture Multiple Social Contexts”. In: *CoRR* abs/1905.02138 (2019). arXiv: 1905.02138. URL: http://arxiv.org/abs/1905.02138.

[107] Gunther Eysenbach. “Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet”. In: *J Med Internet Res* 11.1 (Mar. 2009), e11. ISSN: 1438-8871. DOI: 10.2196/jmir.1157. URL: http://www.ncbi.nlm.nih.gov/pubmed/19329408.

[108] Pegah Falinouss. “Stock Trend Prediction using News Events”. In: *Masters thesis* (2007).

[109] FAO. *The State of Food Security and Nutrition in the World 2019: Safeguarding Against Economic Slowdowns and Downturns*. The State of Food Security and Nutrition in the World. United Nations, 2019. ISBN: 9789210043632. URL: https://books.google.com/books?id=wlrjDwAAQBAJ.

[110] Amir Feder et al. “CausaLM: Causal Model Explanation Through Counterfactual Language Models”. In: *CoRR* abs/2005.13407 (2020). arXiv: 2005.13407. URL: https://arxiv.org/abs/2005.13407.

[111] Michael Feldman et al. “Certifying and Removing Disparate Impact”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. Sydney, NSW, Australia: ACM, 2015, pp. 259–268. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783311. URL: http://doi.acm.org/10.1145/2783258.2783311.

[112] Anjalie Field and Yulia Tsvetkov. *Unsupervised Discovery of Implicit Gender Bias*. 2020. arXiv: 2004.08361 [cs.CL].

[113] Ronald Aylmer Fisher. “Statistical methods for research workers”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 66–70.

[114] George Forman. “An extensive empirical study of feature selection metrics for text classification”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1289–1305.

[115] Paula Fortuna, Juan Soler, and Leo Wanner. “Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets”. English. In:

Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 6786–6794. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.838.

[116] Dean Foster and Rakesh Vohra. "An Economic Argument for Affirmative Action". In: *Rationality and Society* 4.2 (1992), pp. 176–188. DOI: 10.1177/1043463192004002004. eprint: https://doi.org/10.1177/1043463192004002004. URL: https://doi.org/10.1177/1043463192004002004.

[117] Chris Fraley and Adrian E. Raftery. "Model-based clustering, discriminant analysis, and density estimation". In: *Journal of the American statistical Association* 97 (2002), pp. 611–631.

[118] Clark C. Freifeld et al. "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports". In: *Journal of the American Medical Informatics Association* 15.2 (2008), pp. 150–157. DOI: 10.1197/jamia.M2544. URL: http://dx.doi.org/10.1197/jamia.M2544.

[119] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[120] Daniel Fried, Tamara Polajnar, and Stephen Clark. "Low-Rank Tensors for Verbs in Compositional Distributional Semantics". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 731–736. DOI: 10.3115/v1/P15-2120. URL: https://www.aclweb.org/anthology/P15-2120.

[121] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[122] Roy Frostig, Matthew James Johnson, and Chris Leary. "Compiling machine learning programs via high-level tracing". In: *Systems for Machine Learning* (2018), pp. 23–24.

[123] Lisheng Fu et al. "Domain Adaptation for Relation Extraction with Domain Adversarial Neural Network". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 425–429. URL: https://www.aclweb.org/anthology/I17-2072.

[124] David Galles and Judea Pearl. "An axiomatic characterization of causal counterfactuals". In: *Foundations of Science* 3.1 (1998), pp. 151–182.

[125] Juan L. Gamella and Christina Heinze-Deml. "Active Invariant Causal Prediction: Experiment Selection through Stability". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 15464–15475. URL: https://proceedings.neurips.cc/paper/2020/file/b197ffdef2ddc3308584dce7afa3661b-Paper.pdf.

[126] Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. "Modeling Document-level Causal Structures for Event Causal Relation Identification". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1808–1817. DOI: 10.18653/v1/N19-1179. URL: https://www.aclweb.org/anthology/N19-1179.

[127] Qiaozi Gao et al. "What Action Causes This? Towards Naive Physical Action-Effect Prediction". In: *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 934–945. DOI: 10.18653/v1/P18-1086. URL: https://www.aclweb.org/anthology/P18-1086.

[128] Matt Gardner et al. "Evaluating NLP Models via Contrast Sets". In: *CoRR* abs/2004.02709 (2020). arXiv: 2004.02709. URL: https://arxiv.org/abs/2004.02709.

[129] Sahaj Garg et al. "Counterfactual Fairness in Text Classification through Robustness". In: *CoRR* abs/1809.10610 (2018). arXiv: 1809.10610. URL: http://arxiv.org/abs/1809.10610.

[130] Sahaj Garg et al. "Counterfactual Fairness in Text Classification through Robustness". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 219–226. ISBN: 9781450363242. DOI: 10.1145/3306618.3317950. URL: https://doi.org/10.1145/3306618.3317950.

[131] Saurabh Garg et al. "A unified view of label shift estimation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3290–3300.

[132] Lisa Gaudette and Nathalie Japkowicz. "Evaluation Methods for Ordinal Classification". In: *Advances in Artifical Intelligence* (2009), pp. 207–210.

[133] T. Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[134] Timnit Gebru et al. "Datasheets for Datasets". In: *CoRR* abs/1803.09010 (2018). arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010.

[135] Andrew Gelman and Jennifer Hill. "Causal inference using regression on the treatment variable". In: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006, pp. 167–198. DOI: 10.1017/CBO9780511790942.012.

[136] E. D. Gennatas et al. “Expert-augmented machine learning”. In: *PNAS* 117.9 (2020), pp. 4571–4577.

[137] Andreas Gerhardus and Jakob Runge. “High-recall causal discovery for autocorrelated time series with latent confounders”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12615–12625. URL: https://proceedings.neurips.cc/paper/2020/file/94e70705efae423efda1088614128d0b-Paper.pdf.

[138] Charles J Geyer. “Markov chain Monte Carlo maximum likelihood”. In: (1991).

[139] Abbas Ghaddar et al. “End-to-End Self-Debiasing Framework for Robust NLU Training”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1923–1929. DOI: 10.18653/v1/2021.findings-acl.168. URL: https://aclanthology.org/2021.findings-acl.168.

[140] A. Ghazimatin et al. “PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems”. In: *WSDM*. Houston, TX, USA, 2020, pp. 196–204.

[141] Ona de Gibert et al. *Hate Speech Dataset from a White Supremacy Forum*. 2018. arXiv: 1809.04444 [cs.CL].

[142] Gretchen L. Gierach et al. “Hypertension, Menopause, and Coronary Artery Disease Risk in the Women’s Ischemia Syndrome Evaluation (WISE) Study”. In: *Journal of the American College of Cardiology* 47.3_Supplement (2006), S50–S58. DOI: 10.1016/j.jacc.2005.02.099.

[143] Nils Petter Gledistch et al. “Armed Conflict 1946-2001: A New Dataset”. In: *Journal of Peace Research* 39.5 (2002), pp. 615–637. DOI: 10.1177/0022343302039005007. eprint: https://doi.org/10.1177/0022343302039005007. URL: https://doi.org/10.1177/0022343302039005007.

[144] P Godfrey, R Shipley, and J Gryz. “Algorithms and Analyses for Maximal Vector Computation”. In: *The VLDB Journal* 16.1 (Jan. 2007), pp. 5–28. ISSN: 1066-8888. DOI: 10.1007/s00778-006-0029-7. URL: http://dx.doi.org/10.1007/s00778-006-0029-7.

[145] Georg Goergen et al. “First report of outbreaks of the fall armyworm Spodoptera frugiperda (JE Smith)(Lepidoptera, Noctuidae), a new alien invasive pest in West and Central Africa”. In: *PloS one* 11.10 (2016), e0165632.

[146] Janaına Gomide et al. “Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter”. In: *Proceedings of the 3rd International Web Science Conference*. WebSci ’11. Koblenz, Germany: ACM, 2011, 3:1–3:8. ISBN: 978-1-4503-0855-7. DOI: 10.1145/2527031.2527049. URL: http://doi.acm.org/10.1145/2527031.2527049.

[147] Hila Gonen and Yoav Goldberg. *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them*. 2019. arXiv: 1903.03862 [cs.CL].

[148] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[149] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *International Conference on Learning Representations*. 2015.

[150] A. Gordo and F. Perronnin. “Asymmetric Distances for Binary Embeddings”. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR ’11. USA: IEEE Computer Society, 2011, pp. 729–736. ISBN: 9781457703942. DOI: 10.1109/CVPR.2011.5995505. URL: https://doi.org/10.1109/CVPR.2011.5995505.

[151] Diana Gordon and Marie desJardins. “Evaluation and Selection of Biases in Machine Learning”. In: *Machine Learning* 20 (July 1995), pp. 5–22. DOI: 10.1007/BF00993472.

[152] S. Gowal et al. “On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models”. In: (2018). arXiv: 1810.12715.

[153] Tanya Goyal and Greg Durrett. "Embedding Time Expressions for Deep Temporal Ordering Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4400–4406. DOI: 10.18653/v1/P19-1433. URL: https://www.aclweb.org/anthology/P19-1433.

[154] C. W. J. Granger. "Essays in Econometrics". In: ed. by Eric Ghysels, Norman R. Swanson, and Mark W. Watson. Cambridge, MA, USA: Harvard University Press, 2001. Chap. Investigating Causal Relations by Econometric Models and Cross-spectral Methods, pp. 31–47. ISBN: 0-521-79697-0. URL: http://dl.acm.org/citation.cfm?id=781840.781842.

[155] Clive Granger. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–38. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:37:y:1969:i:3:p:424-38.

[156] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[157] Clive WJ Granger, Bwo-Nung Huangb, and Chin-Wei Yang. "A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu". In: *The Quarterly Review of Economics and Finance* 40.3 (2000), pp. 337–354.

[158] N Grgic-Hlaca. "The Case for Process Fairness in Learning : Feature Selection for Fair Decision Making". In: 2016.

[159] N Grgic-Hlaca et al. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction". In: WWW '18. 2018, pp. 903–912. DOI: 10.1145/3178876.3186138. URL: https://doi.org/10.1145/3178876.3186138.

[160] Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France:

European Language Resources Association, May 2020, pp. 1780–1790. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.220.

[161] Aditya Grover and Jure Leskovec. *node2vec: Scalable Feature Learning for Networks*. 2016. arXiv: 1607.00653 [cs.SI].

[162] A. Gunawardana and C. Meek. "A unified approach to building hybrid recommender systems". In: *RecSys*. 2009, pp. 117–124.

[163] Sachin Gupta and Pradeep K. Chintagunta. "On using demographic variables to determine segment membership in logit mixture models". In: *Journal of Marketing Research* (1994), pp. 128–136.

[164] G. Hadash, O. S. Shalom, and R. Osadchy. "Rank and rate: multi-task learning for recommender systems". In: *RecSys*. 2018, pp. 451–454.

[165] Michael Hagenau, Michael Liebmann, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context-capturing features". In: *Decision Support Systems* 55.3 (2013), pp. 685–697. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2013.02.006. URL: http://www.sciencedirect.com/science/article/pii/S0167923613000651.

[166] Rowan Hall Maudslay et al. "It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5266–5274. DOI: 10.18653/v1/D19-1530. URL: https://www.aclweb.org/anthology/D19-1530.

[167] William L. Hamilton, Rex Ying, and Jure Leskovec. "Representation Learning on Graphs: Methods and Applications". In: *CoRR* abs/1709.05584 (2017). arXiv: 1709.05584. URL: http://arxiv.org/abs/1709.05584.

[168] M Hardt, E Price, and N Srebro. “Equality of Opportunity in Supervised Learning”. In: *CoRR* abs/1610.02413 (2016). arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413.

[169] Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning”. In: *CoRR* abs/1610.02413 (2016). arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413.

[170] Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of opportunity in supervised learning”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 3323–3331.

[171] F. Maxwell Harper and Joseph A. Konstan. “The MovieLens Datasets: History and Context”. In: *ACM Trans. Interact. Intell. Syst.* 5.4 (Dec. 2015). ISSN: 2160-6455.

[172] Chikara Hashimoto et al. “Generating Event Causality Hypotheses Through Semantic Relations”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 2396–2403. ISBN: 0-262-51129-0. URL: http://dl.acm.org/citation.cfm?id=2886521.2886654.

[173] Taha Hassan. “Trust and trustworthiness in social recommender systems”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 529–532.

[174] X. He et al. “Adversarial Personalized Ranking for Recommendation”. In: *SIGIR*. 2018, pp. 355–364.

[175] T. Heckelei and H. Wolff. “Estimation of constrained optimisation models for agricultural supply analysis based on generalised maximum entropy”. In: *European review of agricultural economics* 30.1 (2003), pp. 27–50.

[176] H Heidari et al. “Fairness Behind a Veil of Ignorance: Welfare Analysis for Automated Decision Making”. In: abs/1806.04959 (’18). arXiv: 1806.04959. URL: http://arxiv.org/abs/1806.04959.

[177] Stefan Heindorf et al. “CauseNet: Towards a Causality Graph Extracted from the Web”. In: *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*. CIKM ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 3023–3030. ISBN: 9781450368599. DOI: 10.1145/3340531.3412763. URL: https://doi.org/10.1145/3340531.3412763.

[178] Stefan Heindorf et al. “CauseNet: Towards a Causality Graph Extracted from the Web”. In: *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*. CIKM ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 3023–3030. ISBN: 9781450368599. DOI: 10.1145/3340531.3412763. URL: https://doi.org/10.1145/3340531.3412763.

[179] Iris Hendrickx et al. “ILK: Machine learning of semantic relations with shallow features and almost no data”. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 187–190. URL: https://www.aclweb.org/anthology/S07-1039.

[180] Miguel A Hernán and James M Robins. *Causal inference*. 2010.

[181] J. R. Hershey et al. “Deep clustering: Discriminative embeddings for segmentation and separation”. In: *2016 IEEE International Conference on Acoustics*. Speech and Signal Processing (ICASSP, 2016, pp. 31–35.

[182] M. R. Hestenes. “Multiplier and gradient methods”. In: *Journal of Optimization Theory and Applications* 4 (1969), pp. 303–320.

[183] G. Hinton, O. Vinyals, and J. Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[184] Johan Hovold. “Naive Bayes Spam Filtering Using Word-Position-Based Attributes.” In: *CEAS*. 2005, pp. 41–48.

[185] Yu-Lun Hsieh et al. "On the Robustness of Self-Attentive Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1520–1529. DOI: 10.18653/v1/P19-1147. URL: https://www.aclweb.org/anthology/P19-1147.

[186] William Huang, Haokun Liu, and Samuel R. Bowman. "Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data". In: *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 82–87. DOI: 10.18653/v1/2020.insights-1.13. URL: https://aclanthology.org/2020.insights-1.13.

[187] N. J. Hurley. "Robustness of Recommender Systems". In: *RecSys*. 2011, pp. 9–10.

[188] A Ilyas et al. "Adversarial Examples Are Not Bugs, They Are Features". In: *NeurIPS*. Vol. 32. 2019, pp. 125–136.

[189] Guido W Imbens. "Nonparametric estimation of average treatment effects under exogeneity: A review". In: *Review of Economics and statistics* 86.1 (2004), pp. 4–29.

[190] National Center for Immunizations and Division of Viral Diseases Respiratory Diseases. "Middle East Respiratory Syndrome (MERS)". In: (May 2018). URL: https://www.cdc.gov/features/novelcoronavirus/index.html.

[191] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. *First Quora Dataset Release: Question Pairs*. 2017. URL: https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs (visited on 04/03/2019).

[192] Alon Jacovi and Yoav Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://www.aclweb.org/anthology/2020.acl-main.386.

[193] Anil K. Jain. "Data clustering: 50 years beyond k-means". In: *Pattern recognition letters* 31 (2010), pp. 651–666.

[194] Peter Jansen, Mihai Surdeanu, and Peter Clark. "Discourse Complements Lexical Semantics for Non-factoid Answer Reranking". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 977–986. DOI: 10.3115/v1/P14-1092. URL: https://www.aclweb.org/anthology/P14-1092.

[195] Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. *How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks*. 2017. arXiv: 1702.02170 [cs.CL].

[196] Neal Jean et al. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794. ISSN: 0036-8075. DOI: 10.1126/science.aaf7894. URL: https://science.sciencemag.org/content/353/6301/790.

[197] Andrew Jesson et al. "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 11637–11649. URL: https://proceedings.neurips.cc/paper/2020/file/860b37e28ec7ba614f00f9246949561d-Paper.pdf.

[198] Rohan Jha, Charles Lovering, and Ellie Pavlick. "When does data augmentation help generalization in NLP?" In: *CoRR* abs/2004.15012 (2020). arXiv: 2004.15012. URL: https://arxiv.org/abs/2004.15012.

[199] R. Jia et al. "Certified Robustness to Adversarial Word Substitutions". In: *EMNLP-IJCNLP* (2019).

[200] Robin Jia and Percy Liang. "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *CoRR* abs/1707.07328 (2017). arXiv: 1707.07328. URL: http://arxiv.org/abs/1707.07328.

[201] Robin Jia et al. “Certified Robustness to Adversarial Word Substitutions”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2019.

[202] Robin Jia et al. “Certified Robustness to Adversarial Word Substitutions”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4129–4142. DOI: 10.18653/v1/D19-1423. URL: https://www.aclweb.org/anthology/D19-1423.

[203] Yichen Jiang and Mohit Bansal. “Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2726–2736. DOI: 10.18653/v1/P19-1262. URL: https://www.aclweb.org/anthology/P19-1262.

[204] Joshi Kalyani, H. N. Bharathi, and Rao Jyothi. “Stock trend prediction using news sentiment analysis”. In: *CoRR* abs/1607.01958 (2016). arXiv: 1607.01958. URL: http://arxiv.org/abs/1607.01958.

[205] Wagner A. Kamakura. “A least squares procedure for benefit segmentation with conjoint experiments”. In: *Journal of Marketing Research* 25 (1988), pp. 157–67.

[206] Wagner A. Kamakura, Byung-Do Kim, and Jonathan Lee. “Modeling preference and structural heterogeneity in consumer choice”. In: *Marketing Science* 15 (1996), pp. 152–172.

[207] Wagner A. Kamakura and Gary Russell. “A probabilistic choice model for market segmentation and elasticity structure”. In: *Journal of Marketing Research* 26 (1989), pp. 379–390.

[208] Dongyeop Kang et al. “Detecting and Explaining Causes From Text For a Time Series Event”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2758–2767. URL: https://www.aclweb.org/anthology/D17-1292.

[209] W.C. Kang et al. "Learning Multi-Granular Quantized Embeddings for Large-Vocab Categorical Features in Recommender Systems". In: *WWW*. Taipei, Taiwan, 2020, pp. 562–566.

[210] H Kannan, A Kurakin, and I Goodfellow. "Adversarial logit pairing". In: *arXiv:1803.06373* (2018).

[211] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. "End-to-End Bias Mitigation by Modelling Biases in Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8706–8716. DOI: 10.18653/v1/2020.acl-main.769. URL: https://aclanthology.org/2020.acl-main.769.

[212] D. Kaushik, E. Hovy, and Z. C. Lipton. "Learning the difference that makes a difference with counterfactually-augmented data". In: *arXiv:1909.12434* (2019).

[213] Divyansh Kaushik, Eduard H. Hovy, and Zachary C. Lipton. "Learning the Difference that Makes a Difference with Counterfactually-Augmented Data". In: *CoRR* abs/1909.12434 (2019). arXiv: 1909.12434. URL: http://arxiv.org/abs/1909.12434.

[214] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. "Learning The Difference That Makes A Difference With Counterfactually-Augmented Data". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=Sklgs0NFvr.

[215] Divyansh Kaushik et al. "Explaining the Efficacy of Counterfactually Augmented Data". In: *International Conference on Learning Representations (ICLR)* (2021).

[216] Noriaki Kawamae. "Trend analysis model: trend consists of temporal words, topics, and timestamps". In: WSDM '11. 2011, pp. 317–326.

[217] Michael Kearns et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: *ICML*. 2018.

[218] M Kearns et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: *CoRR* abs/1711.05144 (2017).

[219] Katherine Keith, David Jensen, and Brendan O’Connor. “Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5332–5344. DOI: 10.18653/v1/2020.acl-main.474. URL: https://www.aclweb.org/anthology/2020.acl-main.474.

[220] Nitish Shirish Keskar et al. “CTRL: A Conditional Transformer Language Model for Controllable Generation”. In: *CoRR* abs/1909.05858 (2019). arXiv: 1909.05858. URL: http://arxiv.org/abs/1909.05858.

[221] N Kilbertus et al. “Blind Justice: Fairness with Encrypted Sensitive Attributes”. In: ICML. 2018.

[222] Minyoung Kim. “Time-Series Dimensionality Reduction via Granger Causality”. In: *IEEE Signal Processing Letters* 19.10 (2012), pp. 611–614. DOI: 10.1109/LSP.2012.2209641.

[223] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).

[224] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. “Optimization by simulated annealing”. In: *science* 220.4598 (1983), pp. 671–680.

[225] Jon Kleinberg et al. “Prediction Policy Problems”. In: *American Economic Review* 105.5 (2015), pp. 491–95. DOI: 10.1257/aer.p20151023. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20151023.

[226] B. P. Knijnenburg, N. J.M. Reijmer, and M. C. Willemsen. “Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems”. In: *RecSys*. Chicago, Illinois, USA, 2011, pp. 141–148.

[227] Igor Kononenko. “Inductive and Bayesian learning in medical diagnosis”. In: *Applied Artificial Intelligence* 7 (1993), pp. 317–337.

[228] Zornitsa Kozareva. “Cause-effect Relation Learning”. In: *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*. TextGraphs-7 ’12. Jeju, Republic of Korea: Association for Computational Linguistics, 2012, pp. 39–43. URL: http://dl.acm.org/citation.cfm?id=2392954.2392961.

[229] Emmanouil Krasanakis et al. “Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-Aware Classification”. In: *Proceedings of the 2018 World Wide Web Conference*. WWW ’18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 853–862. ISBN: 9781450356398. DOI: 10.1145/3178876.3186133. URL: https://doi.org/10.1145/3178876.3186133.

[230] S. Krichene, M. Gartrell, and C. Calauzènes. “Embedding models for recommendation under contextual constraints”. In: *arXiv* abs/1907.01637 (2019).

[231] Roland Kuhn and Renato De Mori. “Cache-based natural language model for speech recognition”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12 (July 1990), pp. 570–583. DOI: 10.1109/34.56193.

[232] Keita Kurita et al. “Measuring Bias in Contextualized Word Representations”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 166–172. DOI: 10.18653/v1/W19-3823. URL: https://www.aclweb.org/anthology/W19-3823.

[233] Matt J Kusner et al. “Counterfactual Fairness”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017, pp. 4066–4076. URL: https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[234] Matt Kusner et al. “From word embeddings to document distances”. In: *International conference on machine learning* (2015), pp. 957–966.

[235] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. “CASTLE: Regularization via Auxiliary Causal Graph Discovery”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1501–1512. URL: https://proceedings.neurips.cc/paper/2020/file/1068bceb19323fe72b2b344ccf85c254-Paper.pdf.

[236] Carlene MM Lawes et al. “Statin use in COPD patients is associated with a reduction in mortality: a national cohort study”. In: *Primary Care Respiratory Journal* 21.1 (2012), pp. 35–40.

[237] David Lazer et al. “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176 (2014), pp. 1203–1205. ISSN: 0036-8075. DOI: 10.1126/science.1248506. eprint: http://science.sciencemag.org/content/343/6176/1203.full.pdf. URL: http://science.sciencemag.org/content/343/6176/1203.

[238] David Lazer et al. “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176 (2014), pp. 1203–1205. ISSN: 0036-8075. DOI: 10.1126/science.1248506. URL: https://science.sciencemag.org/content/343/6176/1203.

[239] Quoc V. Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents”. In: *CoRR* abs/1405.4053 (2014). arXiv: 1405.4053. URL: http://arxiv.org/abs/1405.4053.

[240] D. G. Lee et al. “Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction”. In: *Journal of Medical Systems* 37.2 (Jan. 2013), p. 9896.

[241] Daniel D. Lee and H. Sebastian Seung. “Algorithms for Non-Negative Matrix Factorization”. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. NIPS’00. Denver, CO: MIT Press, 2000, pp. 535–541.

[242] Kalev Leetaru and Philip A. Schrodt. “GDELT: Global data on events, location, and tone”. In: *ISA Annual Convention* (2013).

[243] EC Lentz et al. “A data-driven approach improves food insecurity crisis prediction”. In: *World Development* 122 (2019), pp. 399–409.

[244] Xiang Lisa Li and Jason Eisner. “Specializing Word Embeddings (for Parsing) by Information Bottleneck”. In: *CoRR* abs/1910.00163 (2019). arXiv: 1910.00163. URL: http://arxiv.org/abs/1910.00163.

[245] Yitong Li, Timothy Baldwin, and Trevor Cohn. “What’s in a Domain? Learning Domain-Robust Text Representations using Adversarial Training”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 474–479. DOI: 10.18653/v1/N18-2076. URL: https://www.aclweb.org/anthology/N18-2076.

[246] Yujia Li et al. *Learning Deep Generative Models of Graphs*. 2018. URL: https://openreview.net/forum?id=Hy1d-ebAb.

[247] Yu Li et al. “Learning Network Embedding with Community Structural Information.” In: *IJCAI*. 2019, pp. 2937–2943.

[248] Z. C. Lipton. “The Mythos of Model Interpretability”. In: (2016). arXiv: 1606.03490.

[249] Z Lipton, J McAuley, and A Chouldechova. "Does mitigating ML's impact disparity require treatment disparity?" In: *NeurIPS 31*. 2018.

[250] Zachary C. Lipton. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." In: *Queue* 16.3 (June 2018), pp. 31–57. ISSN: 1542-7730. DOI: 10.1145/3236386.3241340. URL: https://doi.org/10.1145/3236386.3241340.

[251] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. "Detecting and correcting for label shift with black box predictors". In: *International conference on machine learning*. PMLR. 2018, pp. 3122–3130.

[252] Miaofeng Liu et al. "Reinforced Training Data Selection for Domain Adaptation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1957–1968. DOI: 10.18653/v1/P19-1189. URL: https://www.aclweb.org/anthology/P19-1189.

[253] Weijie Liu et al. "K-BERT: Enabling Language Representation with Knowledge Graph". In: *CoRR* abs/1909.07606 (2019). arXiv: 1909.07606. URL: http://arxiv.org/abs/1909.07606.

[254] Yan Liu et al. "Learning Temporal Causal Graphs for Relational Time-Series Analysis". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 687–694. ISBN: 9781605589077.

[255] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[256] K. Louhichi et al. Modelling Farm-household Livelihoods in Developing Economies Insights from three country case studies using LSMS-ISA data, 2020.

[257] Aurelie C. Lozano et al. “Grouped Graphical Granger Modeling Methods for Temporal Causal Modeling”. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’09. Paris, France: Association for Computing Machinery, 2009, pp. 577–586. ISBN: 9781605584959. DOI: 10.1145/1557019.1557085. URL: https://doi.org/10.1145/1557019.1557085.

[258] Aurelie C Lozano et al. “Spatial-temporal causal modeling for climate change attribution”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 587–596.

[259] Kaiji Lu et al. “Gender Bias in Neural Natural Language Processing”. In: *CoRR* abs/1807.11714 (2018). arXiv: 1807.11714. URL: http://arxiv.org/abs/1807.11714.

[260] Kaiji Lu et al. *Gender Bias in Neural Natural Language Processing*. 2019. arXiv: 1807.11714 [cs.CL].

[261] Zhibin Lu, Pan Du, and Jian-Yun Nie. *VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification*. 2020. arXiv: 2004.05707 [cs.CL].

[262] Chen Luo et al. “Correlating Events with Time Series for Incident Diagnosis”. In: KDD ’14. 2014, pp. 1583–1592. DOI: 10.1145/2623330.2623374.

[263] Xiaofei Ma et al. “Domain Adaptation with BERT-based Domain Classification and Data Selection”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 76–83. DOI: 10.18653/v1/D19-6109. URL: https://www.aclweb.org/anthology/D19-6109.

[264] Douglas L. Maclachlan and Johny K. Johansson. “Market segmentation with multivariate aid”. In: *The Journal of Marketing* (1981), pp. 74–84.

[265] Nishtha Madaan et al. “Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text”. In: *CoRR* abs/2012.04698 (2020). arXiv: 2012.04698. URL: https://arxiv.org/abs/2012.04698.

[266] D Madras et al. “Fairness Through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data”. In: FAT* ’19. 2019.

[267] Asha M. Mahajan et al. “Seasonal and circadian patterns of myocardial infarction by coronary artery disease status and sex in the ACTION Registry-GWTG”. In: *International Journal of Cardiology* 274 (2019), pp. 16–20. ISSN: 0167-5273. DOI: https://doi.org/10.1016/j.ijcard.2018.08.103. URL: https://www.sciencedirect.com/science/article/pii/S0167527318337781.

[268] Valentin Malykh, Varvara Logacheva, and Taras Khakhulin. “Robust Word Vectors: Context-Informed Embeddings for Noisy Texts”. In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 54–63. DOI: 10.18653/v1/W18-6108. URL: https://www.aclweb.org/anthology/W18-6108.

[269] Christopher D Manning, Hinrich Schütze, et al. *Foundations of statistical natural language processing*. Vol. 999. MIT Press, 1999.

[270] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[271] S Mannor, D Peleg, and R Rubinstein. “The Cross Entropy Method for Classification”. In: ICML ’05. 2005.

[272] Huina Mao, Scott Counts, and Johan Bollen. “Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data”. In: *Arxiv preprint* (2011).

[273] Donald Martin-Jr. et al. *Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics*. 2020. arXiv: 2005.07572 [cs.CY].

[274] Ruben Martinez-Cantin. "BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 3735–3739.

[275] N Martinez, M Bertran, and G Sapiro. "Minimax Pareto Fairness: A Multi Objective Perspective". In: 2020.

[276] Andreu Mas-Colell, Michael D Whinston, Jerry R Green, et al. "Chapter 16: Equilibrium and its Basic Welfare Properties". In: *Microeconomic Theory* 1 (1995).

[277] P. Massa and P. Avesani. "Trust-aware recommender systems". In: *RecSys*. 2007.

[278] Daniel Maxwell and Merry Fitzpatrick. "The 2011 Somalia famine: Context, causes, and complications". In: *Global Food Security* 1.1 (2012), pp. 5–12.

[279] Daniel Maxwell et al. "Viewpoint: Determining famine: Multi-dimensional analysis for the twenty-first century". In: *Food Policy* 92 (2020), p. 101832. ISSN: 0306-9192. DOI: https://doi.org/10.1016/j.foodpol.2020.101832. URL: https://www.sciencedirect.com/science/article/pii/S0306919220300166.

[280] Mariusz Maziarz. "A review of the Granger-causality fallacy". In: *The Journal of Philosophical Economics* 8.2 (2015), p. 6. URL: https://EconPapers.repec.org/RePEc:bus:jphile:v:8:y:2015:i:2:n:6.

[281] J. McAuley, R. Pandey, and J. Leskovec. "Inferring networks of substitutable and complementary products". In: *KDD*. 2015, pp. 785–794.

[282] Tom McCoy, Ellie Pavlick, and Tal Linzen. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://www.aclweb.org/anthology/P19-1334.

[283] Geoffrey McLachlan and David Peel. *Finite mixture models*. John & Sons: Wiley, 2004.

[284] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The annals of statistics* (2006), pp. 1436–1462.

[285] John W. Mellor and Sarah Gavian. "Famine: Causes, Prevention, and Relief". In: *Science* 235.4788 (1987), pp. 539–545. ISSN: 0036-8075. DOI: 10.1126/science.235.4788.539. URL: https://science.sciencemag.org/content/235/4788/539.

[286] Facundo Mémoli, Anastasios Sidiropoulos, and Vijay Sridhar. "Quasimetric embeddings and their applications". In: *CoRR* abs/1608.01396 (2016). arXiv: 1608.01396. URL: http://arxiv.org/abs/1608.01396.

[287] Xin Meng, Nancy Qian, and Pierre Yared. "The Institutional Causes of China's Great Famine, 1959-1961". In: *The Review of Economic Studies* 82.4 (293) (2015), pp. 1568–1611. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/43869477.

[288] A K Menon and R C Williamson. "The cost of fairness in binary classification". In: FAT* 2018. 2018.

[289] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: NIPS'13 (2013), pp. 3111–3119.

[290] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[291] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[292] Timothy Miller. "Simplified Neural Unsupervised Domain Adaptation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 414–419. DOI: 10.18653/v1/N19-1039. URL: https://www.aclweb.org/anthology/N19-1039.

[293] Paramita Mirza and Sara Tonelli. "CATENA: CAusal and TEmporal relation extraction from NAtural language texts". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 64–75. URL: http://www.aclweb.org/anthology/C16-1007.

[294] Bhavana Dalvi Mishra et al. *Everything Happens for a Reason: Discovering the Purpose of Actions in Procedural Text*. 2019. arXiv: 1909.04745 [cs.CL].

[295] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *CoRR* abs/1810.03993 (2018). arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993.

[296] A. M. Mohammad et al. "Demographic, clinical and angiographic profile of coronary artery disease in kurdistan region of Iraq". In: *Am J Cardiovasc Dis* 11.1 (2021), pp. 39–45.

[297] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

[298] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[299] Nafise Sadat Moosavi et al. "Improving Robustness by Augmenting Training Sentences with Predicate-Argument Structures". In: *CoRR* abs/2010.12510 (2020). arXiv: 2010.12510. URL: https://arxiv.org/abs/2010.12510.

[300] K. Morik, P. Brockhausen, and T. Joachims. “Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring”. In: *ICML*. 1999, pp. 268–277.

[301] Seithuti P Moshokoa. “On completeness of quasi-pseudometric spaces”. In: *International journal of mathematics and mathematical sciences* 2005.18 (2005), pp. 2933–2943.

[302] R.K Mothilal, A Sharma, and C Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. In: *CoRR* abs/1905.07697 (2019).

[303] Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. “Language understanding for text-based games using deep reinforcement learning”. In: *arXiv preprint arXiv:1506.08941* (2015).

[304] Andrew Y. Ng. “Feature Selection, L₁ vs. L₂ Regularization, and Rotational Invariance”. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 78. ISBN: 1581138385. DOI: 10.1145/1015330.1015435. URL: https://doi.org/10.1145/1015330.1015435.

[305] Andrew Y. Ng et al. *2002*. On spectral clustering: Analysis and an algorithm.

[306] Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. “SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness”. In: *CoRR* abs/2009.10195 (2020). arXiv: 2009.10195. URL: https://arxiv.org/abs/2009.10195.

[307] Jianmo Ni, Jiacheng Li, and Julian McAuley. “Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 188–197. DOI: 10.18653/v1/D19-1018. URL: https://aclanthology.org/D19-1018.

[308] Qiang Ning et al. “Joint Reasoning for Temporal and Causal Relations”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2278–2288. DOI: 10.18653/v1/P18-1212. URL: https://www.aclweb.org/anthology/P18-1212.

[309] Armineh Nourbakhsh et al. “"Breaking" Disasters: Predicting and Characterizing the Global News Value of Natural and Man-made Disasters”. In: *CoRR* abs/1709.02510 (2017). arXiv: 1709.02510. URL: http://arxiv.org/abs/1709.02510.

[310] M P O’Mahony, N J Hurley, and G Silvestre. “Recommender systems: Attack types and strategies”. In: *AAAI*. 2005, pp. 334–339.

[311] M O’Mahony et al. “Collaborative Recommendation: A Robustness Analysis”. In: *ACM Trans. Internet Technol.* (2004).

[312] I. Olkin and F. Pukelsheim. “The distance between two random vectors with given dispersion matrices”. In: *Linear Algebra and its Applications* 48 (1982), pp. 257–263. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(82)90112-4. URL: http://www.sciencedirect.com/science/article/pii/0024379582901124.

[313] Alexandra Olteanu, Onur Varol, and Emre Kiciman. “Distilling the Outcomes of Personal Experiences: A Propensity-Scored Analysis of Social Media”. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’17. Portland, Oregon, USA: Association for Computing Machinery, 2017, pp. 370–386. ISBN: 9781450343350. DOI: 10.1145/2998181.2998353. URL: https://doi.org/10.1145/2998181.2998353.

[314] Alexandra Olteanu, Onur Varol, and Emre Kiciman. “Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media”. In: *Proceedings of The 20th*

ACM Conference on Computer-Supported Cooperative Work and Social Computing. Association for Computing Machinery, Inc., Feb. 2017. ISBN: 978-1-4503-4335-0/17/03.

[315] Yonatan Oren et al. "Distributionally Robust Language Modeling". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4227–4237. DOI: 10.18653/v1/D19-1432. URL: https://www.aclweb.org/anthology/D19-1432.

[316] World Health Organization. "Anatomical Therapeutic Chemical (ATC) Classification System with Defined Daily Doses". In: (2003).

[317] World Health Organization. "Conflict and Infectious Diseases". In: (Mar. 2019). URL: https://www.who.int/tdr/research/social_research/conflict/en/.

[318] World Health Organization. *International classification of diseases*. 1978.

[319] World Health Organization. "WHO MERS Disease Outbreak News". In: (Aug. 2018). URL: http://www.who.int/csr/don/archive/disease/coronavirus_infections/en/.

[320] World Health Organization. "WHO MERS Global Summary and Assessment of Risk, Disease Outbreak News". In: (Aug. 2018). URL: http://www.who.int/csr/disease/coronavirus_infections/risk-assessment-august-2018.pdf.

[321] Malte Ostendorff et al. *Enriching BERT with Knowledge Graph Embeddings for Document Classification*. 2019. arXiv: 1909.08402 [cs.CL].

[322] Mingdong Ou et al. "Asymmetric Transitivity Preserving Graph Embedding". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1105–1114. ISBN: 9781450342322. DOI: 10.1145/2939672.2939751. URL: https://doi.org/10.1145/2939672.2939751.

[323] Yaniv Ovadia et al. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift*. 2019. arXiv: 1906.02530 [stat.ML].

[324] N Papernot et al. "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data". In: *CoRR* abs/1610.05755 (2016). URL: http://arxiv.org/abs/1610.05755.

[325] Vilfredo Pareto. *Manuale di economia politica: con una introduzione alla scienza sociale*. Vol. 13. Società editrice libraria, 1919.

[326] Ji Ho Park, Jamin Shin, and Pascale Fung. "Reducing Gender Bias in Abusive Language Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2799–2804. DOI: 10.18653/v1/D18-1302. URL: https://www.aclweb.org/anthology/D18-1302.

[327] M Pawelczyk, K Broelemann, and G Kasneci. "Learning Model-Agnostic Counterfactual Explanations for Tabular Data". In: WWW '20. 2020.

[328] Judea Pearl. "Causal inference in statistics: An overview". In: *Statistics Surveys* (2009).

[329] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc., 2018. ISBN: 046509760X.

[330] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[331] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. "DeepWalk: Online Learning of Social Representations". In: *CoRR* abs/1403.6652 (2014). arXiv: 1403.6652. URL: http://arxiv.org/abs/1403.6652.

[332] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. "Causal inference on time series using restricted structural equation models". In: *Advances in Neural Information Processing Systems*. 2013, pp. 154–162.

[333] Therése Pettersson and Kristine Eck. "Organized violence, 1989-2017". In: *Journal of Peace Research 55* (2018).

[334] Silviu Pitis, Elliot Creager, and Animesh Garg. *Counterfactual Data Augmentation using Locally Factored Dynamics*. 2020. arXiv: 2007.02863 [cs.LG].

[335] G Pleiss et al. "On Fairness and Calibration". In: *CoRR* abs/1709.02012 (2017). arXiv: 1709.02012. URL: http://arxiv.org/abs/1709.02012.

[336] Michael JD Powell. "The BOBYQA algorithm for bound constrained optimization without derivatives". In: *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge* 26 (2009).

[337] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. "Perturbation Sensitivity Analysis to Detect Unintended Model Biases". In: *CoRR* abs/1910.04210 (2019). arXiv: 1910.04210. URL: http://arxiv.org/abs/1910.04210.

[338] Sameer Pradhan et al. "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 1–27. URL: https://www.aclweb.org/anthology/W11-1901.

[339] Sameer Pradhan et al. "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1–40. URL: https://www.aclweb.org/anthology/W12-4501.

[340] B. Prakash, A. Jaiswal, and M. M. Shah. "Demographic & angiographic profile of young patients aged 40 year & less undergoing coronary angiography in a tier II city of Eastern India". In: *J Family Med Prim Care* 9.10 (Oct. 2020), pp. 5183–5187.

[341] R. Prashanth. "Finding association between lipid profile and demographic and disease status of patients undergoing coronary angiography: A retrospective study in rural South India". In: *JOURNAL OF INDIAN COLLEGE OF CARDIOLOGY* 11.2 (2021), pp. 62–65. DOI: 10.4103/JICC.JICC_56_20. eprint: https://www.joicc.org/article.asp?issn=1561-8811;year=2021;volume=11;issue=2;spage=62;epage=65;aulast=Prashanth;t=6. URL: https://www.joicc.org/article.asp?issn=1561-8811;year=2021;volume=11;issue=2;spage=62;epage=65;aulast=Prashanth;t=6.

[342] Robert J. Prill et al. "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges". In: *PLOS ONE* 5.2 (Feb. 2010), pp. 1–18. DOI: 10.1371/journal.pone.0009202. URL: https://doi.org/10.1371/journal.pone.0009202.

[343] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. "Combating Adversarial Misspellings with Robust Word Recognition". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5582–5591. DOI: 10.18653/v1/P19-1561. URL: https://www.aclweb.org/anthology/P19-1561.

[344] C Qin et al. "Adversarial robustness through local linearization". In: *NeurIPS*. 2019.

[345] T Qin et al. "LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval". In: *Inf. Retr.* 13.4 (2010). ISSN: 1386-4564.

[346] Y Qin et al. "Improving Uncertainty Estimates through the Relationship with Adversarial Robustness". In: *ArXiv* 2006.16375 (2020).

[347] Kira Radinsky and Eric Horvitz. "Mining the web to predict future events". In: WSDM '13. ACM. 2013, pp. 255–264.

[348] Tushar Rao and Saket Srivastava. “Analyzing Stock Market Movements Using Twitter Sentiment Analysis”. In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. ASONAM ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 119–123. ISBN: 978-0-7695-4799-2. DOI: 10.1109/ASONAM.2012.30. URL: http://dx.doi.org/10.1109/ASONAM.2012.30.

[349] B Rastegarpanah, K P Gummadi, and M Crovella. “Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems”. In: *WWW*. 2019, pp. 231–239.

[350] Alexander Ratner et al. “Snorkel: Rapid Training Data Creation with Weak Supervision”. In: *CoRR* abs/1711.10160 (2017). arXiv: 1711.10160. URL: http://arxiv.org/abs/1711.10160.

[351] Alexander Ratner et al. “Snorkel: Rapid Training Data Creation with Weak Supervision”. In: *CoRR* abs/1711.10160 (2017). arXiv: 1711.10160. URL: http://arxiv.org/abs/1711.10160.

[352] Martin Ravallion. “Famines and Economics”. In: *Journal of Economic Literature* 35.3 (1997), pp. 1205–1242. ISSN: 00220515. URL: http://www.jstor.org/stable/2729976.

[353] John Rawls. *A theory of justice*. 1971.

[354] Shuhuai Ren et al. “Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1085–1097. DOI: 10.18653/v1/P19-1103. URL: https://www.aclweb.org/anthology/P19-1103.

[355] M.T Ribeiro, S Singh, and C Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *CoRR* abs/1602.04938 (2016).

[356] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Semantically Equivalent Adversarial Rules for Debugging NLP models". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 856–865. DOI: 10.18653/v1/P18-1079. URL: https://www.aclweb.org/anthology/P18-1079.

[357] Marco Tulio Ribeiro et al. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. URL: https://www.aclweb.org/anthology/2020.acl-main.442.

[358] Alberto Riva and Riccardo Bellazzi. "Learning temporal probabilistic causal models from longitudinal data". In: *Artificial Intelligence in Medicine* 8.3 (1996). Temporal Reasoning in Medicine, pp. 217–234. ISSN: 0933-3657. DOI: https://doi.org/10.1016/0933-3657(95)00034-8. URL: https://www.sciencedirect.com/science/article/pii/0933365795000348.

[359] Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for Confounding with Text Matching". In: *American Journal of Political Science* 64.4 (2020), pp. 887–903. DOI: https://doi.org/10.1111/ajps.12526. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12526. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12526.

[360] Lior Rokach and Oded Maimon. "Clustering methods". In: *Data mining and knowledge discovery handbook*. 2005, pp. 321–352.

[361] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[362] Alexis Ross, Ana Marasovic, and Matthew E. Peters. "Explaining NLP Models via Minimal Contrastive Editing (MiCE)". In: *CoRR* abs/2012.13985 (2020). arXiv: 2012.13985. URL: https://arxiv.org/abs/2012.13985.

[363] G N. Rothblum and G Yona. "Probably Approximately Metric-Fair Learning". In: *CoRR* abs/1803.03242 (2018). arXiv: 1803.03242. URL: http://arxiv.org/abs/1803.03242.

[364] Paul K Rubenstein et al. "Causal consistency of structural equation models". In: *arXiv preprint arXiv:1707.00819* (2017).

[365] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[366] Rachel Rudinger et al. *Gender Bias in Coreference Resolution*. 2018. arXiv: 1804.09301 [cs.CL].

[367] Keisuke Sakaguchi et al. *WinoGrande: An Adversarial Winograd Schema Challenge at Scale*. 2019. arXiv: 1907.10641 [cs.CL].

[368] Haji Mohammad Saleem et al. *A Web of Hate: Tackling Hateful Speech in Online Social Spaces*. 2017. arXiv: 1709.10159 [cs.CL].

[369] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis. Wiley Series in Probability and Statistics*. 2000.

[370] E.I Sato et al. "Demographic, clinical, and angiographic data of patients with Takayasu arteritis in Brazil". In: *International Journal of Cardiology* 66 (1998), S67–S70. ISSN: 0167-5273. DOI: https://doi.org/10.1016/S0167-5273(98)00152-1. URL: https://www.sciencedirect.com/science/article/pii/S0167527398001521.

[371] Robert E Schapire. "The boosting approach to machine learning: An overview". In: *Nonlinear estimation and classification*. Springer, 2003, pp. 149–171.

[372] Bernhard Schölkopf. “Causality for Machine Learning”. In: *CoRR* abs/1911.10500 (2019). arXiv: 1911.10500. URL: http://arxiv.org/abs/1911.10500.

[373] A D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: FAT* ’19. 2019.

[374] A. Sen. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford India paperbacks. Oxford University Press, 1982. URL: https://books.google.com/books?id=b8TRPQAACAAJ.

[375] Uri Shalit, Fredrik D. Johansson, and David Sontag. *Estimating individual treatment effect: generalization bounds and algorithms*. 2017. arXiv: 1606.03976 [stat.ML].

[376] J Shang et al. “Pre-training of Graph Augmented Transformers for Medication Recommendation”. In: *CoRR* abs/1906.00346 (2019).

[377] Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. “Degree based Classification of Harmful Speech using Twitter Data”. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 106–112. URL: https://www.aclweb.org/anthology/W18-4413.

[378] Rebecca Sharp et al. *Creating Causal Embeddings for Question Answering with Minimal Supervision*. 2016. arXiv: 1609.08097 [cs.CL].

[379] Rebecca Sharp et al. “Spinning Straw into Gold: Using Free Text to Train Monolingual Alignment Models for Non-factoid Question Answering”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 231–237. DOI: 10.3115/v1/N15-1025. URL: https://www.aclweb.org/anthology/N15-1025.

[380] John Shawe-Taylor, Martin Anthony, and Norman Biggs. “Bounding Sample Size with the Vapnik-Chervonenkis Dimension”. In: 42 (Feb. 1993), pp. 65–73.

[381] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on pattern analysis and machine intelligence* 22 (2000), pp. 888–905.

[382] Y. Shynkevich et al. "Stock price prediction based on stock-specific and sub-industry-specific news articles". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. July 2015, pp. 1–8. DOI: 10.1109/IJCNN.2015.7280517.

[383] Brian E. Simmons et al. "Coronary artery disease in blacks of lower socioeconomic status: Angiographic findings from the Cook County Hospital Heart Disease Registry". In: *American Heart Journal* 116.1, Part 1 (1988), pp. 90–97. ISSN: 0002-8703. DOI: https://doi.org/10.1016/0002-8703(88)90254-2. URL: https://www.sciencedirect.com/science/article/pii/0002870388902542.

[384] Ashudeep Singh and Thorsten Joachims. "Fairness of Exposure in Rankings". In: *KDD*. Ed. by Yike Guo and Faisal Farooq. ACM, 2018, pp. 2219–2228.

[385] A Sinha, D. F. Gleich, and K Ramani. "Deconvolving Feedback Loops in Recommender Systems". In: *CoRR* abs/1703.01049 (2017).

[386] Nathaniel Smilowitz et al. "Mortality of Myocardial Infarction by Sex, Age, and Obstructive Coronary Artery Disease Status in the ACTION Registry–GWTG (Acute Coronary Treatment and Intervention Outcomes Network Registry–Get With the Guidelines)". In: *Circulation: Cardiovascular Quality and Outcomes* 10 (Dec. 2017), e003443. DOI: 10.1161/CIRCOUTCOMES.116.003443.

[387] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[388] Richard Socher et al. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural*

Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. URL: https://www.aclweb.org/anthology/D13-1170.

[389] T Speicher et al. "Potential for Discrimination in Online Targeted Advertising". In: *ACM FAccT*. 2018.

[390] Peter Spirtes. "An Anytime Algorithm for Causal Inference." In: *AISTATS*. 2001.

[391] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Vol. 81. Jan. 1993. ISBN: 978-1-4612-7650-0. DOI: 10.1007/978-1-4612-2748-9.

[392] Harald Steck. "Embarrassingly Shallow Autoencoders for Sparse Data". In: *CoRR* abs/1905.03375 (2019). arXiv: 1905.03375. URL: http://arxiv.org/abs/1905.03375.

[393] Jiankai Sun et al. "ATP: Directed Graph Embedding with Asymmetric Transitivity Preservation". In: *CoRR* abs/1811.00839 (2018). arXiv: 1811.00839. URL: http://arxiv.org/abs/1811.00839.

[394] Swabha Swayamdipta et al. "Syntactic Scaffolds for Semantic Structures". In: *Empirical Methods in Natural Language Processing* (2018), pp. 3772–3782. DOI: 10.18653/v1/D18-1412. URL: https://www.aclweb.org/anthology/D18-1412.

[395] Robert H Swendsen and Jian-Sheng Wang. "Replica Monte Carlo simulation of spin-glasses". In: *Physical review letters* 57.21 (1986), p. 2607.

[396] Saifuddin Syed et al. "Non-reversible parallel tempering: A scalable highly parallel MCMC scheme". In: *arXiv preprint arXiv:1905.02939* (2019).

[397] A Taha and A Hadi. "Anomaly Detection Methods for Categorical Data: A Review". In: *ACM Comput. Surv.* 52.2 (May 2019). ISSN: 0360-0300.

[398] Jian Tang et al. "LINE: Large-scale Information Network Embedding". In: *CoRR* abs/1503.03578 (2015). arXiv: 1503.03578. URL: http://arxiv.org/abs/1503.03578.

[399] A. Tank et al. “Neural Granger Causality for Nonlinear Time Series”. In: *ArXiv e-prints* (Feb. 2018). arXiv: 1802.05842 [stat.ML].

[400] Alex Tank et al. *Neural Granger Causality for Nonlinear Time Series*. 2018. arXiv: 1802.05842 [stat.ML].

[401] Zenna Tavares et al. “Predicate exchange: Inference with declarative knowledge”. In: *International Conference on Machine Learning*. 2019, pp. 6186–6195.

[402] Joshua B Tenenbaum, Vin De Silva, and John C Langford. “A global geometric framework for nonlinear dimensionality reduction”. In: *science* 290.5500 (2000), pp. 2319–2323.

[403] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. *Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision*. 2020. arXiv: 2004.09034 [cs.CV].

[404] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. In: *Journal of the Royal Statistical Society, Series B* 58 (1994), pp. 267–288.

[405] Brendon Towle and Clark N. Quinn. “Knowledge Based Recommender Systems Using Explicit User Models”. In: 2000.

[406] Florian Tramer et al. “Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 9561–9571. URL: http://proceedings.mlr.press/v119/tramer20a.html.

[407] Cunchao Tu et al. “Max-Margin Deepwalk: Discriminative Learning of Network Representation”. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI’16. New York, New York, USA: AAAI Press, 2016, pp. 3889–3895. ISBN: 9781577357704.

[408] UNICEF. “Statistics and Monitoring: Country Statistics”. In: (Aug. 2015). URL: https://www.unicef.org/statistics/index_countrystats.html.

[409] Supreme Court of the United States. “Griggs v. Duke Power Co.” In: 401 U.S. 424 (1971).

[410] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. “Towards Debiasing NLU Models from Unknown Biases”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7597–7610. DOI: 10.18653/v1/2020.emnlp-main.613. URL: https://aclanthology.org/2020.emnlp-main.613.

[411] Carmen K Vaca et al. “A time-based collective factorization for topic discovery and monitoring in news”. In: WWW ’14. 2014, pp. 527–538.

[412] J. Vacca and H. Rosenthal. *A Local Law in relation to automated decision systems used by agencies*. 2018. URL: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

[413] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Second. Springer, Nov. 1999. ISBN: 0387987800.

[414] Victor Veitch, Dhanya Sridhar, and David M. Blei. *Adapting Text Embeddings for Causal Inference*. 2020. arXiv: 1905.12741 [cs.LG].

[415] Petar Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[416] P Velickovic et al. “Graph Attention Networks”. In: *ArXiv* abs/1710.10903 (2018).

[417] Ishan Verma, Lipika Dey, and Hardik Meisheri. “Detecting, Quantifying and Accessing Impact of News Events on Indian Stock Indices”. In: *Proceedings of the International Conference on Web Intelligence*. WI ’17. Leipzig, Germany: ACM, 2017, pp. 550–557. ISBN: 978-

1-4503-4951-2. DOI: 10.1145/3106426.3106482. URL: http://doi.acm.org/10.1145/3106426.3106482.

[418] Sahil Verma, John Dickerson, and Keegan Hines. *Counterfactual Explanations for Machine Learning: A Review*. 2020. arXiv: 2010.10596 [cs.LG].

[419] J Villena-Román et al. "Hybrid approach combining machine learning and a rule-based expert system for text categorization". In: *Twenty-Fourth International FLAIRS Conference*. 2011.

[420] S.V.N. Vishwanathan et al. "Graph Kernels". In: *Journal of Machine Learning Research* 11.40 (2010), pp. 1201–1242. URL: http://jmlr.org/papers/v11/vishwanathan10a.html.

[421] Ellen Viste, Diriba Korecha, and Asgeir Sorteberg. "Recent drought and precipitation tendencies in Ethiopia". In: *Theoretical and Applied Climatology* 112.3 (2013), pp. 535–551.

[422] Alex Wang et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://www.aclweb.org/anthology/W18-5446.

[423] Dieter Wang et al. "Stochastic modeling of food insecurity". In: *World Bank Policy Research Working Papers* (2020).

[424] J Wang and J Caverlee. "Recurrent recommendation with local coherence". In: WSDM. 2019.

[425] Tianlu Wang et al. "CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5141–5146. DOI: 10.18653/v1/2020.emnlp-main.417. URL: https://www.aclweb.org/anthology/2020.emnlp-main.417.

[426] Wei Wang et al. “Growing pains for global monitoring of societal events”. In: *Science* 353.6307 (2016), pp. 1502–1503. ISSN: 0036-8075. DOI: 10.1126/science.aaf6758. URL: https://science.sciencemag.org/content/353/6307/1502.

[427] Xiao Wang et al. “Community preserving network embedding.” In: *AAAI*. Vol. 17. 10.5555. 2017, pp. 3298239–3298270.

[428] Yixin Wang and David M. Blei. “The Blessings of Multiple Causes.” In: vol. abs/1805.06826. 2018. URL: http://dblp.uni-trier.de/db/journals/corr/corr1805.html#abs-1805-06826.

[429] Yu Wang, Eugene Agichtein, and Michele Benzi. “Tm-lda: efficient online modeling of latent topic transitions in social media”. In: KDD ’12. ACM. 2012, pp. 123–131.

[430] Z Wang et al. “A path-constrained framework for discriminating substitutable and complementary products in e-commerce”. In: *WWW*. 2018.

[431] Zeerak Waseem and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”. In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. DOI: 10.18653/v1/N16-2013. URL: https://www.aclweb.org/anthology/N16-2013.

[432] S. Webber and C. Prouse. “The new gold standard: The rise of randomized control trials and experimental development”. In: *Economic Geography* 94.2 (2018), pp. 166–187.

[433] Kellie Webster et al. *Measuring and Reducing Gendered Correlations in Pre-trained Models*. 2020. arXiv: 2010.06032 [cs.CL].

[434] Michel Wedel and Wayne S. DeSarbo. “A review of recent developments in latent class regression models”. In: *Advanced methods of marketing research* (1994), pp. 352–388.

[435] Michel Wedel and Cor Kistemaker. “Consumer benefit segmentation using clusterwise linear regression”. In: *International Journal of Research in Marketing* 6 (1989), pp. 45–59.

[436] Michel Wedel and Jan-Benedict E. M. Steenkamp. “A fuzzy clusterwise regression approach to benefit segmentation”. In: *International Journal of Research in Marketing* 6 (1989), pp. 241–258.

[437] Michael Wellman and Max Henrion. “Explaining ‘explaining away’”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15 (Apr. 1993), pp. 287–292.

[438] John Wieting and Kevin Gimpel. “ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 451–462. DOI: 10.18653/v1/P18-1042. URL: https://www.aclweb.org/anthology/P18-1042.

[439] A. B. Wilcox and G. Hripcsak. “The role of domain knowledge in automating medical text report classification.” In: *JAMIA* (2003).

[440] B Woodworth et al. “Learning Non-Discriminatory Predictors”. In: *COLT*. 2017.

[441] World Bank. “World Development Report 2021: Data For Better Lives”. In: (2021).

[442] Tongshuang Wu et al. “Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021.

[443] Xing Wu et al. *Conditional BERT Contextual Augmentation*. 2018. arXiv: 1812.06705 [cs.CL].

[444] Boyi Xie et al. “Semantic Frames to Predict Stock Price Movement”. In: *ACL (1)*. The Association for Computer Linguistics, 2013, pp. 873–883.

[445] C Xie et al. “Feature denoising for improving adversarial robustness”. In: *ICCV*. 2019.

[446] Junyuan Xie, Ross Girshick, and Ali Farhadi. “Unsupervised deep embedding for clustering analysis”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML’16)*. JMLR.org, 2016, pp. 478–487.

[447] Zhipeng Xie and Feiteng Mu. “Distributed Representation of Words in Cause and Effect Spaces”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 7330–7337. DOI: 10.1609/aaai.v33i01.33017330. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4720.

[448] D Xin et al. “Folding: Why Good Models Sometimes Make Spurious Recommendations”. In: RecSys ’17. 2017.

[449] Depeng Xu et al. “FairGAN: Fairness-aware Generative Adversarial Networks”. In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 570–575. DOI: 10.1109/BigData.2018.8622525.

[450] Rui Xu and Donald Wunsch. “Survey of clustering algorithms”. In: *IEEE Transactions on neural networks* 16 (2005), pp. 645–678.

[451] Yao-Yuan Yang et al. “Adversarial Robustness Through Local Lipschitzness”. In: *CoRR* abs/2003.02460 (2020). URL: https://arxiv.org/abs/2003.02460.

[452] Kai Hang Yiu et al. “Age- and gender-specific differences in the prognostic value of CT coronary angiography”. In: *Heart* 98.3 (2012), pp. 232–237. ISSN: 1355-6037. DOI: 10.1136/heartjnl-2011-300038. eprint: https://heart.bmj.com/content/98/3/232.full.pdf. URL: https://heart.bmj.com/content/98/3/232.

[453] Benny Yong and Livia Owen. “Dynamical transmission model of MERS-CoV in two areas”. In: *AIP Conference Proceedings* 1716.1 (2016), p. 020010. DOI: 10.1063/1.4942993. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.4942993. URL: https://aip.scitation.org/doi/abs/10.1063/1.4942993.

[454] Jiaxuan You et al. “GraphRNN: A Deep Generative Model for Graphs”. In: *CoRR* abs/1802.08773 (2018). arXiv: 1802.08773. URL: http://arxiv.org/abs/1802.08773.

[455] Victor L. Yu and Lawrence C. Madoff. “ProMED-mail: An Early Warning System for Emerging Diseases”. In: *Clinical Infectious Diseases* 39.2 (2004), pp. 227–232. DOI: 10.1086/422003. eprint: /oup/backfile/content_public/journal/cid/39/2/10.1086_422003/3/39-2-227.pdf. URL: http://dx.doi.org/10.1086/422003.

[456] M B Zafar et al. “From Parity to Preference-based Notions of Fairness in Classification”. In: 2017.

[457] Jiehang Zeng et al. *Certified Robustness to Text Adversarial Attacks by Randomized [MASK]*. 2021. arXiv: 2105.03743 [cs.CL].

[458] Guanhua Zhang et al. “Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4134–4145. DOI: 10.18653/v1/2020.acl-main.380. URL: https://www.aclweb.org/anthology/2020.acl-main.380.

[459] H Zhang et al. “Theoretically Principled Trade-off between Robustness and Accuracy”. In: CoRR/1901.08573 (2019).

[460] J. Zhang and Elias Bareinboim. “Fairness in Decision-Making - The Causal Explanation Formula”. In: *AAAI*. 2018.

[461] Yuan Zhang, Jason Baldridge, and Luheng He. *PAWS: Paraphrase Adversaries from Word Scrambling*. 2019. arXiv: 1904.01130 [cs.CL].

[462] Z Zhang et al. “Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion”. In: *ICCV, 2018*. 2018, pp. 1372–1380.

[463] Jieyu Zhao et al. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: https://www.aclweb.org/anthology/N18-2003.

[464] Jieyu Zhao et al. “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2979–2989. DOI: 10.18653/v1/D17-1323. URL: https://www.aclweb.org/anthology/D17-1323.

[465] J Zhao et al. “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”. In: *EMNLP*. 2017.

[466] Q Zhao et al. “Categorical-Attributes-Based Item Classification for Recommender Systems”. In: RecSys ’18. 2018.

[467] Sendong Zhao et al. “Constructing and Embedding Abstract Event Causality Networks from Text Snippets”. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM ’17. Cambridge, United Kingdom: ACM, 2017, pp. 335–344. ISBN: 978-1-4503-4675-7. DOI: 10.1145/3018661.3018707. URL: http://doi.acm.org/10.1145/3018661.3018707.

[468] Zhengli Zhao, Dheeru Dua, and Sameer Singh. “Generating Natural Adversarial Examples”. In: *ICLR*. 2018.

[469] Shi Zhong and Joydeep Ghosh. “A unified framework for model-based clustering”. In: *The Journal of Machine Learning Research* 4 (2003), pp. 1001–1037.

[470] Chang Zhou et al. *Scalable Graph Embedding for Asymmetric Proximity*. 2017. URL: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14696/14500.

[471] G Zhou et al. “Deep Interest Network for Click-Through Rate Prediction”. In: ACM SIGKDD ’18. 2018. ISBN: 9781450355520.

[472] Yichao Zhou et al. “Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4904–4913. DOI: 10.18653/v1/D19-1496. URL: https://www.aclweb.org/anthology/D19-1496.

[473] Shijie Zhu et al. *Adversarial Directed Graph Embedding*. 2020. arXiv: 2008.03667 [cs.SI].

[474] Ran Zmigrod et al. “Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1651–1661. DOI: 10.18653/v1/P19-1161. URL: https://www.aclweb.org/anthology/P19-1161.

[475] Ran Zmigrod et al. “Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology”. In: *CoRR* abs/1906.04571 (2019). arXiv: 1906.04571. URL: http://arxiv.org/abs/1906.04571.

[476] R Zmigrod et al. “Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology”. In: *arXiv preprint arXiv:1906.04571* (2019).