Enhancing Neural Recommender Models through Domain-Specific Concordance

Ananth Balashankar
New York University, Google AI
New York, USA
ananth@nyu.edu

Alex Beutel
Google AI
New York, USA

Lakshminarayanan Subramanian
New York University
New York, USA

ABSTRACT

Recommender models trained on historical observational data alone can be brittle when domain experts subject them to counterfactual evaluation. In many domains, experts can articulate common, high-level mappings or rules between categories of inputs (user's history) and categories of outputs (preferred recommendations). One challenge is to determine how to train recommender models to adhere to these rules. In this work, we introduce the goal of *domain-specific concordance*: the expectation that a recommender model follow a set of expert-defined categorical rules. We propose a regularization-based approach that optimizes for robustness on rule-based input perturbations. To test the effectiveness of this method, we apply it in a medication recommender model over diagnosis-medicine categories, and in movie and music recommender models, on rules over categories based on movie tags and song genres. We demonstrate that we can increase the category-based robustness distance by up to 126% without degrading accuracy, but rather increasing it by up to 12% compared to baseline models in the popular MIMIC-III, MovieLens-20M and Last.fm Million Song datasets.

CCS CONCEPTS

• **Information systems → Recommender systems**.

ACM Reference Format:
Ananth Balashankar, Alex Beutel, and Lakshminarayanan Subramanian. 2021. Enhancing Neural Recommender Models through Domain-Specific Concordance. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441784

1 INTRODUCTION

Black box neural recommender models trained on only observed historical data can make costly errors, which limit their widespread deployment in scenarios that require domain knowledge [39, 52, 65]. Domain experts in these scenarios are particularly skeptical as black box recommender models often contradict rules derived from domain knowledge that have 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 [56]. For example, in the music recommendation domain, users could potentially expect that the genre of their recommendations would not change, if they swapped one of the songs in their history with another song from the same genre, and in the health domain, doctors may expect recommendations for medications to follow domain-specific rules.

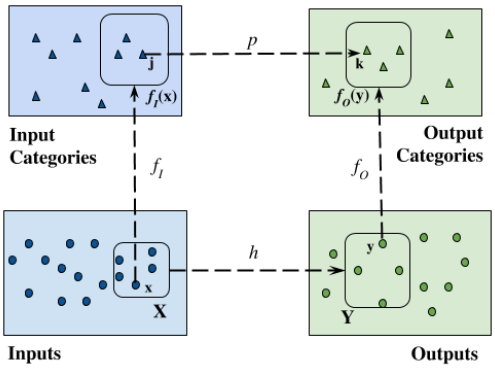

Figure 1: Category-based Rules in Recommenders

We propose a definition of *domain-specific concordance* to reflect this expectation from domain experts that recommender models follow domain specific rules over input and output categories. As shown in Figure 1, for a recommender model h, which maps a subset of inputs X (historical items) to a subset of outputs Y (preferred recommendations), we consider the setting where category mapping functions f_I, f_O map individual input items $x \in X$ and output items $y \in Y$ to their corresponding set of categories $f_I(x), f_O(y)$. The domain specific rule $p(j) = k$ captures the expectation, that if an input x mapped to a category j exists in X, h should recommend an output y in category k, such that $j \in f_I(x)$ and $k \in f_O(y)$.

With domain specific concordance, we expect that the domain-specific rules $p(j) = k$ are generally obeyed by h, but at the same time, we know in personalized applications that these rules may not apply to every possible input or situation. Therefore, to safely make use of these rules beyond the historical data, we optimize for local robustness in the model h, such that if an example (X, Y) matches a domain-specific rule $p(j) = k$ such that $\exists x \in X, \exists y \in Y, j \in f_I(x) \wedge k \in f_O(y)$, the model should also obey that rule by recommending an output in category k for any perturbation x' on the item x that retains the category j, i.e $j \in f_I(x')$. We seek to minimize any *output-category* misclassifications over such *within-category input perturbations*. This framework of robustness in recommender systems to within-category input perturbations to

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441784

align with domain expert-defined rules, and the method of incorporating counterfactual domain specific rule-based data augmentation into training are our main contributions.

Our framing of domain-specific concordance builds on existing robustness research, which answers the question, "How do the model's predictions change under small perturbations to inputs?" In classifiers, this type of trust has been developed through enabling counterfactual explanations [15, 38, 43] and improving robustness in output predictions when inputs have imperceptible and label-invariant perturbations [26, 29, 70]. However, in recommender systems, 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. On the other hand, strictly enforcing fine-grained behaviors in recommenders such as individual user-interaction [30], trust modeling [35] can be hard to achieve and further exacerbated by cold-start problems [2]. Rather, while drawing on this body of work, we leverage domain-specific mappings between categories when generating the counterfactual perturbations. Through our framework, we incorporate the expectations of domain experts by using their rules to generate within-category counterfactual inputs and optimizing neural recommender models to avoid categorical mistakes when presented with them. Hence, we demonstrate the generality and value of our approach by instantiating it over coarsely-defined categories for state-of-the-art recommender models in the domains of movies, music and medicine.

Content Recommendations Domain Concordance: In content (movie and music) recommendations, we employ our framework to the hypothesis that users select content by genre, and as such we optimize for robustness to within-genre perturbations. For example, if a user with a history of watching "animation" movies wanted to watch next another animation movie, then they would still likely have wanted to watch an animation movie, even if a movie (e.g. Toy Story) in their history was replaced with another movie (Toy Story II), both from the animation category. Through this approach, we improved overall accuracy on the original held-out test data by 0.03-2.3%. Further, we evaluate robustness as the distance between the categories of perturbed input items that brings about a category change in the mapped predicted output (see Section 4.3 for a formal description). The improvement in accuracy is driven by an increase in robustness distance by 101-126% as compared to state-of-the-art models on the MovieLens and Million Song datasets. This further confirms the value in enhancing neural recommender models learning through domain specific category mappings in addition to optimizing accuracy on observational data.

Medical Recommendations Domain Concordance: We also test this approach on the problem framing that doctors generally expect that for a patient with a particular category of skin disease (e.g., Dermatophytosis), there is a corresponding category of medications (e.g., Antimycotics) that should typically be prescribed. If our counterfactual patient's attributes changed in only *which* kind of Dermatophytosis they have, we may expect that the category of medication recommended would not change, particularly if doctors have previously confirmed that they prescribe Antimycotics to patients with Dermatophytosis. Incorporating this domain-specific concordance improves overall F1 score by 12.2% on the original test data. Further, this improvement is driven by increase in robustness distance by 75.6% as compared to state-of-the-art baseline models [50]. This is significant as we optimize for recommenders to not alter the category of medication for minor category-invariant perturbations in diagnostic codes as expected by doctor-validated rules. While our contribution is not application specific, and particularly for the medical domain we believe that a medicine recommender would still need significant doctor review, we believe experiments across these domains together provide strong evidence for the usefulness and effectiveness of our proposed technique.

To summarize, our key contributions include:

- **Framework of Domain-Specific Concordance for Recommenders:** We demonstrate how to align a model with domain-specific rules through within-category input perturbations.
- **Optimization through Robustness:** We offer a methodology based on robust within-category regularization that improves adherence to the domain-specific rules
- **Empirical evidence** that our method improves both domain-specific categorical concordance and overall accuracy across recommendation tasks in three domains.

2 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 [14, 58]. Tools to guide the underlying deep learning model through interactive feedback [5] and inductive logic [63] that increases diversity and aligns the model's predictions to expert knowledge have been proposed in the medical domain [37]. Applying data mining to extract association rules using Bayesian methods between input and output categories are also well studied [33], 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 [13]. Our work closely relates to the "task related latent dimensions of interpretability". Here, we care about the hypothesis of local interpretability[49], with incomplete coverage of domain expertise [67]. 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 [6, 8]. However, such adversarial robustness may either increase [25] or decrease [66] 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 [64] and in natural language processing [23], robustness over word substitutions with synonyms are desired [45]. 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 [34, 47]. 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) [59, 65]. Robust models that explicitly guard against multiple attack models [24] like profile injection [10], noisy ratings [42] and implicit issues like outliers [55], data not missing at random [32] have been proposed. Our definition is complementary to prior work in robust recommender models which propose simpler models like decision trees [31], fairness guarantees to avoid unintended bias [3, 4, 9, 51], temporal coherence to avoid catastrophic forgetting [59], defence against adversarial attacks of imperceptible changes [7, 21], and uncertainty based model calibration [65]. 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 [36]. Such substitutable items have been inferred through browsing patterns like "users who viewed X bought Y" and co-purchasing logs [60]. Prior work incorporating categories through hierarchical autoencoders [12], multi-tasking [68], categorical embeddings [27] 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.

3 PROBLEM FORMULATION

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

3.1 Notations

As illustrated in Figure 1, 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 [17] 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: $d_c(k, k')$.

3.2 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 [40]. 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 [41], 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 [61], 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 [61] as priors, we expect that one of the 3 medicines in category "Antimycotics" k would *likely* still be recommended.

3.3 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 3.1. **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:

$$\delta_j(X) = \{x' \cup X \setminus x | x \in X,\ x' \notin X,\ j \in f_I(x),\ \in f_I(x')\} \quad (1)$$

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 3.2. **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 human 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.

4 METHODS

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

4.1 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 4.1. **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

$$\mathcal{L}_v(D_p) = \mathbb{E}_{(X,Y)\in D_p} \mathop{\mathbb{E}}_{\substack{(j,k):p(j)=k \\ X'\sim\delta_j(X)}} \mathbb{I}(k \notin \bigcup_{y'\in h(X')} f_O(y')) \quad (2)$$

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.

$$\mathcal{L}_c(D) = \mathbb{E}_{(X,Y)\in D} \mathcal{L}(h(X), Y) \quad (3)$$

Attempting to write a loss similar to (2), but on the actual counterfactual outputs Y', is difficult as we essentially do not observe them [44] and the changes are not imperceptible. However, by focusing on higher-level categories in (2), 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* [62] 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}:

$$h_{RA} = argmin_h(\alpha\mathcal{L}_v(D_p) + (1-\alpha)\mathcal{L}_c(D)) \quad (4)$$

4.2 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 4.2. **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

$$\mathcal{L}_r(X, X', y) = max(0, max_{y'\in\mathcal{Y}_k}(z_{(X,y)} - z_{(X',y')})) \quad (5)$$

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)**

$$\mathcal{L}_{ar}(D_p) = \mathop{\mathbb{E}}_{\substack{(X,Y)\in D_p,(j,k):p(j)=k \\ X'\sim\delta_j(X), y\sim Y\cap\mathcal{Y}_k}} \mathcal{L}_r(X, X', y) \quad (6)$$

Our approach is related to multiple lines of prior work. For example, interval bounded propagation [16] 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 [22], logit pairing [28] and multi-task modeling [18] 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 1.

Algorithm 1 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

4.3 Metrics

To build the neural recommender models that follow domain rules, we regularize the model such that within-category loss (6) is minimized. We evaluate improvement in robustness using the following distance metric between inputs.

Definition 4.3. **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 $O(X')$.

$$O(X') = \{f_O(y') : \forall y' \in h(X')\} \tag{7}$$

$$S_k(X) = \{x' | X' = x' \cup X \setminus x \wedge x \in X \wedge k \notin O(X')\} \tag{8}$$

$$d_{robust} = \mathbb{E}_{(X,Y)\in D_p}[\min_{\substack{j\in f_I(x), j'\in f_I(x') \\ x\in X, p(j)=k, x'\in S_k(X)}} (d_c(j, j'))] \tag{9}$$

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) [46] 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.

5 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 1, for the domains and rules we consider (Table 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 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 ***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. 2) and ***Baseline* RA-WCR**, which uses Rule-based Augmentation and Within-Category Regularization (Eq. 6). We set $\alpha = 0.2$ after cross-validation.

5.1 Medication Recommendation

We follow the MIMIC-III medication recommendation task as per [50], and the domain specific mappings p are obtained from [61] 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

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 Node Distance
MovieLens	Movie	Movie	Movie Tag	Movie Tag	Identity	Tag Score Difference
Last.fm	Song	Song	Genre, artist type, era	Genre, artist type, era	Identity	Hamming Distance

Table 2: Instantiations of recommender systems into our hybrid framework

pairs to generate perturbations in our existing dataset as shown in Algorithm 1.

5.1.1 Baseline. We use the current state-of-the-art for the medication recommendation task on MIMIC-III dataset as the *Baseline* - G-BERT [50]. 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 [11]. The graph embeddings are learnt using the Graph Attention technique [57], so as to learn hierarchical embeddings for each of the diagnostic and medication codes.

5.2 Movie Recommendation

In the MovieLens dataset [19], 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 1. 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 [20] or polarizing content with feedback loops [48]. 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.

5.2.1 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) [69].

5.3 Music Recommendation

The music recommendation task is taken up on the Million Song dataset from Last.fm [54]. 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

	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 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

DBPedia [1] 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 1. 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⟩.

5.3.1 Baseline. The baseline used is the current state-of-the-art, EASE, which uses shallow autoencoders [53] over the user history. By enforcing that the diagonal of the weight matrix to be zero, to avoid collapse to the trivial identity function, they learn the weights that capture the similarity between songs.

6 EVALUATION

In this section, we evaluate our methodology on all three domains and five model structures from Section 5. For each domain, we study the impact of our method along multiple dimensions to confirm our hypothesis of whether it can improve accuracy (§6.1) and within-category concordance (§6.2). We further perform fine-grained evaluations to understand the source of the changes in accuracy and robustness by coverage, types of rules and popularity (§6.3). 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.

6.1 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 3, 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*

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 4: Our regularized version of DIN with Dice [69] 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 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.

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 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.

6.2 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 4.3*, 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 6, 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 6), 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 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.

6.3 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 2, 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 2 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 accuracy and robustness* as shown in Figure 3. 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

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 MSD) d_{robust} (Hamming Distance)
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 6: Our method considerably increases the mean robustness distance (± standard deviation in brackets - see Def. 4.3) in medication, movie and song domains.

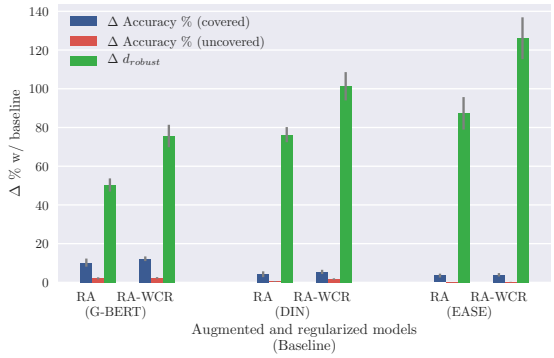

Figure 2: Our method improves robustness (bars are mean, with error bars showing one standard deviation) without degrading accuracy, and improves accuracy the most for subset of data covered by the domain specific mappings.

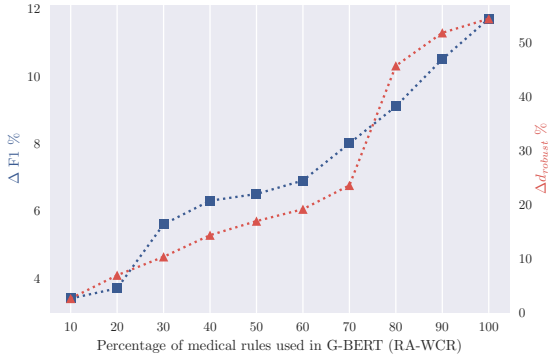

Figure 3: Our G-BERT (RA-WCR) model steadily improves F1 score and robustness distance as and when new medical rules are used to augment the dataset.

versus rules which are not. We bucketize the categorical mappings into five quintiles based on the NMI score in Figure 4, 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 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

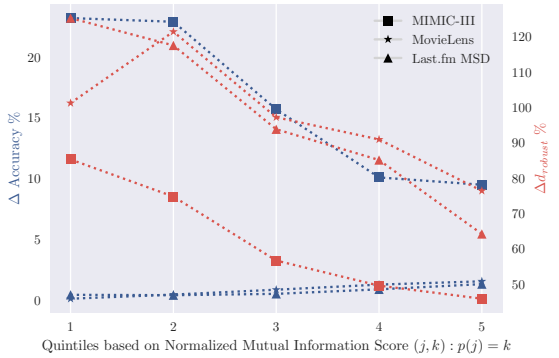

Figure 4: Our RA-WCR approach demonstrates more gain in sliced accuracy and robustness when augmentation is done through rules which have lower normalized mutual information score in the observed data across 3 domains

see in Figure 4, 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 unmodified 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 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.

REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *ISWC'07/ASWC'07* (Busan, Korea). 722–735.

[2] L. Bernardi, J. Kamps, J. Kiseleva, and M. JI Müller. 2015. The continuous cold start problem in e-commerce recommender systems. *arXiv:1508.01177* (2015).

[3] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. *KDD* (2019), 2212–2220.

[4] A. J. Biega, K. P. Gummadi, and G. Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*. ACM, 405–414.

[5] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *CHI*.

[6] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S Liang. 2019. Unlabeled data improves adversarial robustness. In *NeurIPS*. 11192–11203.

[7] K. Christakopoulou and A. Banerjee. 2019. Adversarial Attacks on an Oblivious Recommender. In *RecSys* (Copenhagen, Denmark). 322–330.

[8] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918* (2019).

[9] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogin, and T. Di Noia. 2019. Recommender Systems Fairness Evaluation via Generalized Cross Entropy. arXiv:1908.06708

[10] Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. A. Merra. 2020. How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models. In *SIGIR* (Virtual Event, China). 951–960.

[11] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[12] Y. Ding, J. Liu, and D. Wang. 2018. Deep Feature Fusion over Multi-Field Categorical Data for Rating Prediction. In *AICCC* (Tokyo, Japan). 16–22.

[13] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[14] E. D. Gennatas, J. H. Friedman, L. H. Ungar, R. Pirracchio, E. Eaton, L. G. Reichmann, Y. Interian, J. M. Luna, C. B. Simone, A. Auerbach, E. Delgado, M. J. van der Laan, T. D. Solberg, and G. Valdes. 2020. Expert-augmented machine learning. *PNAS* 117, 9 (2020), 4571–4577.

[15] A. Ghazimatin, O. Balalau, R. Saha Roy, and G. Weikum. 2020. PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems. In *WSDM* (Houston, TX, USA). 196–204.

[16] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. A. Mann, and P. Kohli. 2018. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. (2018). arXiv:1810.12715

[17] A. Gunawardana and C. Meek. 2009. A unified approach to building hybrid recommender systems. In *RecSys*. 117–124.

[18] G. Hadash, O. S. Shalom, and R. Osadchy. 2018. Rank and rate: multi-task learning for recommender systems. In *RecSys*. 451–454.

[19] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages.

[20] Taha Hassan. 2019. Trust and trustworthiness in social recommender systems. In *Companion Proceedings of The 2019 World Wide Web Conference*. 529–532.

[21] X. He, Z. He, X. Du, and T.S. Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *SIGIR*. 355–364.

[22] G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[23] Y.L. Hsieh, M. Cheng, D.C. Juan, W. Wei, W.L. Hsu, and C.J. Hsieh. 2019. On the Robustness of Self-Attentive Models. In *ACL*. Florence, Italy, 1520–1529.

[24] N. J. Hurley. 2011. Robustness of Recommender Systems. In *RecSys*. 9–10.

[25] A Ilyas, S Santurkar, D Tsipras, L Engstrom, B Tran, and A Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *NeurIPS*, Vol. 32. 125–136.

[26] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. 2019. Certified Robustness to Adversarial Word Substitutions. *EMNLP-IJCNLP* (2019).

[27] W.C. Kang, D. Z. Cheng, T. Chen, X. Yi, D. Lin, L. Hong, and E. H. Chi. 2020. Learning Multi-Granular Quantized Embeddings for Large-Vocab Categorical Features in Recommender Systems. In *WWW* (Taipei, Taiwan). 562–566.

[28] H Kannan, A Kurakin, and I Goodfellow. 2018. Adversarial logit pairing. *arXiv:1803.06373* (2018).

[29] D. Kaushik, E. Hovy, and Z. C. Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv:1909.12434* (2019).

[30] B. P. Knijnenburg, N. J.M. Reijmer, and M. C. Willemsen. 2011. Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. In *RecSys* (Chicago, Illinois, USA). 141–148.

[31] Igor Kononenko. 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence* 7 (1993), 317–337.

[32] S. Krichene, M. Gartrell, and C. Calauzènes. 2019. Embedding models for recommendation under contextual constraints. *arXiv* abs/1907.01637 (2019).

[33] D. G. Lee, K. S. Ryu, M. Bashir, J.W. Bae, and K. H. Ryu. 2013. Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction. *Journal of Medical Systems* 37, 2 (15 Jan 2013), 9896.

[34] Z. C. Lipton. 2016. The Mythos of Model Interpretability. (2016). arXiv:1606.03490

[35] P. Massa and P. Avesani. [n.d.]. Trust-aware recommender systems. In *RecSys*.

[36] J. McAuley, R. Pandey, and J. Leskovec. 2015. Inferring networks of substitutable and complementary products. In *KDD*. 785–794.

[37] K. Morik, P. Brockhausen, and T. Joachims. 1999. Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring. In *ICML*. 268–277.

[38] R.K Mothilal, A Sharma, and C Tan. 2019. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *CoRR* abs/1905.07697 (2019).

[39] M P O'Mahony, N J Hurley, and G Silvestre. 2005. Recommender systems: Attack types and strategies. In *AAAI*. 334–339.

[40] World Health Organization. 1978. *International classification of diseases*.

[41] World Health Organization. 2003. Anatomical Therapeutic Chemical (ATC) Classification System with Defined Daily Doses. (2003).

[42] M O'Mahony, N Hurley, N Kushmerick, and G Silvestre. 2004. Collaborative Recommendation: A Robustness Analysis. *ACM Trans. Internet Technol.* (2004).

[43] M Pawelczyk, K Broelemann, and G Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data *(WWW '20)*.

[44] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect* (1st ed.). Basic Books, Inc., USA.

[45] C Qin, J Martens, S Gowal, D Krishnan, K Dvijotham, A Fawzi, S De, R Stanforth, and P Kohli. 2019. Adversarial robustness through local linearization. In *NeurIPS*.

[46] T Qin, T Liu, J Xu, and H Li. 2010. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Inf. Retr.* 13, 4 (2010).

[47] Y Qin, X Wang, A Beutel, and E.H Chi. 2020. Improving Uncertainty Estimates through the Relationship with Adversarial Robustness. *ArXiv* 2006.16375 (2020).

[48] B Rastegarpanah, K P Gummadi, and M Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *WWW*. 231–239.

[49] M.T Ribeiro, S Singh, and C Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016).

[50] J Shang, T Ma, C Xiao, and J Sun. 2019. Pre-training of Graph Augmented Transformers for Medication Recommendation. *CoRR* abs/1906.00346 (2019).

[51] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*, Yike Guo and Faisal Farooq (Eds.). ACM, 2219–2228.

[52] A Sinha, D. F. Gleich, and K Ramani. 2017. Deconvolving Feedback Loops in Recommender Systems. *CoRR* abs/1703.01049 (2017).

[53] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. *CoRR* abs/1905.03375 (2019). arXiv:1905.03375 http://arxiv.org/abs/1905.03375

[54] B Whitman T Bertin-Mahieux, D. P.W. Ellis and P Lamere. 2011. The Million Song Dataset. *ISMIR* (2011).

[55] A Taha and A Hadi. 2019. Anomaly Detection Methods for Categorical Data: A Review. *ACM Comput. Surv.* 52, 2, Article 38 (May 2019).

[56] Brendon Towle and Clark N. Quinn. 2000. Knowledge Based Recommender Systems Using Explicit User Models.

[57] P Velickovic, G Cucurull, A Casanova, A Romero, P Liò, and Y Bengio. 2018. Graph Attention Networks. *ArXiv* abs/1710.10903 (2018).

[58] J Villena-Román, S Collada-Pérez, S Lana-Serrano, and J.C González-Cristóbal. 2011. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *Twenty-Fourth International FLAIRS Conference*.

[59] J Wang and J Caverlee. 2019. Recurrent recommendation with local coherence *(WSDM)*.

[60] Z Wang, Z Jiang, Z Ren, J Tang, and D Yin. 2018. A path-constrained framework for discriminating substitutable and complementary products in e-commerce. In *WWW*.

[61] Filzmoser P Gyimesi M. Weisser A, Endel G. 2008. ATC -> ICD – evaluating the reliability of prognoses for ICD-10 diagnoses derived from the ATC-Code of prescriptions. *BMC Health Serv Res. 2008* (2008).

[62] Michael Wellman and Max Henrion. 1993. Explaining 'explaining away'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15 (04 1993), 287–292.

[63] A. B. Wilcox and G. Hripcsak. 2003. The role of domain knowledge in automating medical text report classification. *JAMIA* (2003).

[64] C Xie, Y Wu, L Maaten, A Yuille, and K He. 2019. Feature denoising for improving adversarial robustness. In *ICCV*.

[65] D Xin, N Mayoraz, H Pham, K Lakshmanan, and J Anderson. 2017. Folding: Why Good Models Sometimes Make Spurious Recommendations *(RecSys '17)*.

[66] H Zhang, Y Yu, J Jiao, E Xing, L Ghaoui, and M Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. CoRR/1901.08573 (2019).

[67] Z Zhang, C Xie, J Wang, L Xie, and A Yuille. 2018. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *ICCV, 2018*. 1372–1380.

[68] Q Zhao, J Chen, M Chen, S Jain, A Beutel, F Belletti, and E. Chi. 2018. Categorical-Attributes-Based Item Classification for Recommender Systems *(RecSys '18)*.

[69] G Zhou, X Zhu, C Song, Y Fan, H Zhu, X Ma, Y Yan, J Jin, H Li, and K Gai. 2018. Deep Interest Network for Click-Through Rate Prediction *(ACM SIGKDD '18)*.

[70] R Zmigrod, S Mielke, H Wallach, and R Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571* (2019).