Learning Causally Disentangled Representations for Fair Personality Detection

Yangfu Zhu[1], **Meiling Li**[2], **Yuting Wei**[2], **Di Liu**[2], **Yuqing Li**[2], **Bin Wu**[2*]
[1] College of Information Engineering, Capital Normal University, Beijing, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
zhuyangfu@cnu.edu.cn, {meilinglee,yuting_wei,liudi,liyuqing,wubin}@bupt.edu.cn

Abstract

Personality detection aims to identify the personality traits implied in social posts. Existing methods mainly focus on learning the mapping between user-generated posts and personality trait labels but inevitably suffer from potential harm caused by individual bias, as these posts are written by authors from different backgrounds. Learning such spurious associations between posts and traits may lead to the formation of stereotypes, ultimately restricting the detection of personality in different kind of individual. To tackle the issue, we first investigate individual bias in personality detection from the causality perspective. We propose an Interventional Personality Detection Network (IPDN) to learn implicit confounders in user-generated posts and exploit the true causal effect to train the detection model. Specifically, our IPDN disentangled the causal and biased features behind user-generated posts, and then the biased features are accumulatively clustered as confounder prototypes as the training iterations increase. In parallel, the reconstruction network is reused to approximate backdoor adjustment on raw posts, ensuring that traits see each confounder equally before detection. Extensive experiments conducted on three real-world datasets demonstrate that our IPDN outperforms state-of-the-art methods in personality detection.

1 Introduction

Personality refers to an individual's psychological constitution, including unique aspects of cognition, emotions, attitudes, and values. With the blossoming of social media, users generate massive posts daily that reveal their psychological activities, providing new opportunities to infer personality traits automatically [Fang *et al.*, 2023]. Successfully inferring traits from such content can enhance massive applications, such as recommendation system [Yang *et al.*, 2022], dialogue system [Chawla *et al.*, 2023], and human-computer interaction design [Chien *et al.*, 2022].

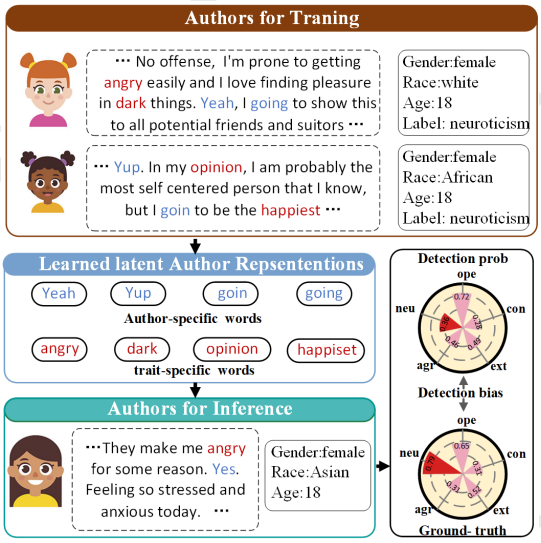


Figure 1: An illustration of the racial bias in personality detection. In the training data, a group White and African American authors exhibit neurotic traits, with their posts containing both words related to neurotic traits and words related to their racial backgrounds. When given test posts with similar traits but generated by Asian authors, the model inevitably infers the wrong personality states.

In previous research works, Deep Neural Networks (DNNs) are the default choice because they can reveal the most predictive hidden patterns from posts [Lynn *et al.*, 2020; Yang *et al.*, 2021b; Zhu *et al.*, 2022; Yang *et al.*, 2023b]. Recently, Large Language Models (LLMs) have been rapidly applied in personality detection, such as psychological questionnaires based chain of thought [Yang *et al.*, 2023c] and large model-based data augmentation [Hu *et al.*, 2024]. Despite impressive improvements in personality detection, these well-trained DNNs may "cheat" us by inadvertently capturing and even amplifying unintended individual biases through author-generated content. As an example, Figure 1 illustrates how racial bias confounds the detection models. To further verify detection modal is not independent of the human individual attributes, we conducted an experimental investigation with unbalanced demographic attributes on the Pandora dataset. As shown in Table 1, the unbalanced group (Group I) always obtains a higher rate of false positives and false neg-

*Corresponding authors.

atives. This demonstrates that although the text does not explicitly contain any individual attribute, well-trained models do incorrectly learn the spurious correlations between posts and traits.

While some statistical studies have explored the relationship between personality and demographics [Gjurković *et al.*, 2021; Kerz *et al.*, 2022], the research on eliminating individual confounders to improve the fairness of personality detection is still a wilderness. To achieve this goal, we face the following challenges: **(1)** Due to privacy concerns, people are reluctant to share personal information on the internet, making it labor-intensive and time-consuming to obtain user-generated posts labeled with individual attributes. **(2)** Even with access to most attributes information, some unknown types of human biases are difficult to predefine in advance.

To solve the above challenges, we attempted to improve the personality detection model by applying causal intervention. we propose a novel **I**nterventional **P**ersonality **D**etection **N**etwork (IPDN) to remove the confounding effect of individual attributes C through backdoor adjustment $P(Y|do(X))$ instead of the conventional likelihood $P(Y|X)$ estimated by traditional DNNs-based methods. Specifically, unlike approximating the intervention via predefined confounders, we turn to learning implicit confounders in user-generated posts to achieve confounder-agnostic general debiasing. Our IPDN starts to decompose the causal feature and biased feature by disentanglement and reconstruction module. Then, a confounder builder cumulatively clusters bias features and approximates confounder prototypes as the iteration of training. Finally, the reconstruction network are reused to approximate backdoor adjustment on raw posts, ensuring that traits see each confounder equally before detection. In summary, our main contributions are as follows:

- To the best of our knowledge, this is the first effort to take individual biases into account for personality detection from the causal perspective. By considering individual attributes as confounding factors, we aim to look for spurious correlations in the model and mitigate unfair detection.
- Based on the causal theory of backdoor adjustment, we propose a novel interventional personality detection network (IPDN) that automatically learns confounders from post and cuts off the spurious correlations between posts and traits through deconfounded training with intervention.
- Extensive experiments demonstrate the effectiveness of IPDN and the fairness of personality detection is improved to a certain extent.

2 Related Work

2.1 Personality Detection

Early efforts heavily relied on psycholinguistic statistics features to detect personality, such as Linguistic Inquiry and Word Count (LIWC) [Tausczik and Pennebaker, 2010] Mairesse [Mehta *et al.*, 2020], and MRC [Coltheart, 1981], considering the influence of personality traits on language use patterns. However, statistical analysis has limitations in effectively capturing the semantics of posts. With the rapid advancement of Deep Neural Networks (DNNs), a line of DNNs approaches like CNNs [Xue *et al.*, 2018], LSTMs [Tandera *et al.*, 2017] and Transformer [Yang *et al.*, 2021a; Zhu *et al.*, 2022] have demonstrated impressive results in this field. Meanwhile, another line of research focus on the structure of user-generated posts, such as hierarchical structure model (SN+Attn) [Lynn *et al.*, 2020], tripartite graph networks (TrigNet) [Yang *et al.*, 2021b], graph contrastive learning (CGTN) [Zhu *et al.*, 2022], and dynamic graph networks (D-DGCN) [Yang *et al.*, 2023b]. Recently, Large Language Models (LLMs) have been rapidly applied in personality detection, The PsyCoT model uses psychological questionnaires as Chain of Thought to simulate personality tests through multiple rounds of dialogue [Yang *et al.*, 2023c]. TAE model tried to distill LLM knowledge to enhance the small model for personality detection [Hu *et al.*, 2024]. However, the above methods mainly focus on obtaining a meaningful representation of user-generated posts, overlooking the influence of the inherent individual attributes of authors. Certainly, some researchers have investigated and confirmed the relationships between demographic variables and personality traits based on psychological theories, the de-biased personality detection remains unexplored to our knowledge.

Traits	Demographic	False Positive (%)		False Negative (%)	
		Group I	Group II	Group I	Group II
I / E	Gender	68.09	47.63	44.45	38.46
I / E	Age	63.71	53.60	47.68	36.65

Table 1: Results of the MLP that trained on two demographic controlled groups. Group I set demographic attributes to 5:1 and group II is 1:1. The Introversion and Extraversion traits are balanced in both groups.

2.2 Causal Inference

Causal inference is developed to estimate causal effect with covariate shift. Benefiting from the great potential of the causal tool to provide unbiased estimation solutions, it has been widely applied to diverse fields, such as recommendation [Wei *et al.*, 2022], emotion recognition [Yang *et al.*, 2023a] natural language inference [Feder *et al.*, 2022] and pretrained language models [Zhou *et al.*, 2023]. The most relevant work to us is the debiased text classification. CORSAIR [Qian *et al.*, 2021] imagines the counterfactual docu'ment to distill and mitigate the label bias and keyword biases in text. CCD [Chen *et al.*, 2023] applies Normalized Weighted Geometric Mean (NWGM) [Xu *et al.*, 2015] approximate causal intervention, removing the psycholinguistic bias in fake news detection. In this work, we discover human bias in personality detection and propose a novel causal intervention-based debiasing framework to enhance detection performance.

3 Methodology

3.1 Problem Definition

Given a set of posts $P = \{p_1, p_2 ... p_n\}$ from a user, the goal of personality detection is to predict the categories of t-dimensional personality traits, denoted as $\mathrm{Y} = \{y_1, y_2, y_t\}$.

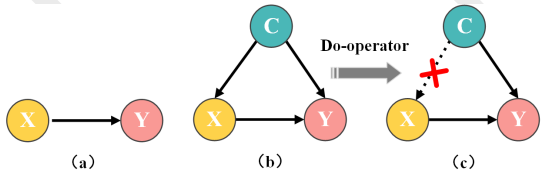


Figure 2: The Structural Causal Models (SCM) of the personality detection process. (a) The SCM of traditional methods. (b) The SCM that considers individual attribute as confounder. (c) The SCM after the causal intervention.

The personality traits are categorized into the Myers-Briggs personality inventory (MBTI) [Myers, 1997] and the Big-Five indicators [Digman, 1990]. the MBTI taxonomy includes *Introversion / Extraversion*, *Sensing / Ntuition*, *Thinking / Feeling*, and *Judging / Perceiving*. While the Big-Five taxonomy includes *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*.

3.2 Causal View at Personality Detection

To answer the whys of the individual attribute affects personality detection, we formulate the Structural Causal Models (SCM) [Pearl, 2009] to illustrate causal relations in the detection process. As shown in Figure 2, it is a directed acyclic graph that consists of three variables (nodes): Node X, Y, and C denote the user-generated posts, the predicted probability of the personality traits, and the individual attributes, respectively. The individual attributes include demographic and some unknown attributes, such as cultural background, emotions. The directed edges are the causalities between two variables: cause $\rightarrow$ effect. $X \rightarrow Y$. Traditional methods only consider that the personality traits could be inferred by user-generated posts. $C \rightarrow Y$. Compared to the traditional method, we realistically add a confounder node C to portray the detection process. The author attribute directly affects the final detected probability, which is expected since detectors can learn to associate post with individual information to make unfair decisions about certain groups. $C \rightarrow X$. Similarly, the individual attributes has a causal effect on posts and indirectly affects the detected probability. Such individual information are normally embedded in the user-generated posts, for instance, the wording and tone of the authors affect their posts.

Based on this causal graph, we can observe the presence of confounder C opens a **backdoor path** as show in Figure 2 (b): $X \leftarrow C \rightarrow Y$ introduces spurious correlation between the posts X and personality traits Y by learning the likelihood $P(Y|X)$, thereby degrading the generalization performance. This process is formulated by bayes rule:

$$P(Y|X) = \sum_{c} P(Y|X, C=c)P(C=c|X), \qquad (1)$$

The confounders C introduce the observational bias via $P(C|X)$, because C satisfy the backdoor criterion that all backdoor paths between X and Y are blocked by conditioning on C and C do not consist of any variables that are descendants of X.

3.3 Intervention via Backdoor Adjustment

Our goal is to develop a detection model that is unaffected by unobserved confounders C. One straightforward way is conducting a randomized controlled trial by collecting any attribute of any individual to ensure that traits are balanced across all confounders. In this case, the conditional probability $P(Y|X)$ equals the causal probability $P(Y|do(X))$. However, implementing such solutions may not be practical. First, users are often unwilling to disclose personal information. Furthermore, the complexity of confounders makes constructing a balanced dataset challenging. We turn to a more elegant intervention on the input posts X by blocking the backdoor path between X and Y based on the backdoor adjustment (Figure 2 C). The backdoor adjustment performs a **do-operator** on X, effectively cutting off the causal path from C to X. As a result, the detection model will approximate the causal intervention $P(Y|do(X))$ instead of the spurious association $P(Y|X)$ estimated by traditional methods:

$$P(Y|do(X)) = \sum_{c} P(Y|X, f(X,c))P(c), \qquad (2)$$

where $f(\cdot)$ is nonlinear transformation that incorporate independent variables X and each confounder c. Through causal intervention, we provide a fair opportunity for X to observe each confounder c when detecting trait Y. After that, the spurious correlation from C to X is cut off, allowing the causal effect $X \rightarrow Y$ free from the effect of C.

3.4 IPDN Architecture

Backdoor adjustment requires confounder C could be stratified. Unfortunately, there is no strict and widely accepted definition for the confounder that influence traits, let alone stratifying for them. To end this, we propose an Interventional Personality Detection Network (IPDN) to automatically learn confounders C and mitigate biases. As shown in Figure 3, the IPDN comprises three modules: 1) the feature disentanglement and reconstruction; 2) the confounder builder; 3) the deconfounded training with intervention.

Feature Disentanglement and Reconstruction

Given the post vector u_i for each user, f_c and f_b are applied as the projection function to learn the causal and biased representations:

$$\boldsymbol{e_i^c} = f_c(\boldsymbol{u_i}, \boldsymbol{\theta_c}), \boldsymbol{e_i^b} = f_b(\boldsymbol{u_i}, \boldsymbol{\theta_b}), \qquad (3)$$

where $\boldsymbol{\theta_c}$ and $\boldsymbol{\theta_b}$ are learnable parameters of the two projection function, respectively. Neural network dynamics research has found that biased features are easier to learn than the desired knowledge in the early stages of training [Nam *et al.*, 2020]. The Generalized Cross Entropy (GCE) loss is used [Zhang and Sabuncu, 2018] to amplify bias. Specifically, we forward the biased features $\boldsymbol{e_i^b}$ for detection:

$$\mathrm{y}_b = softmax(f_b(\boldsymbol{u_i}, \boldsymbol{\theta_b})), \qquad (4)$$

$$\mathcal{L}_{gce}(\mathrm{y}_b, \mathrm{y}) = \frac{1 - p_y(\boldsymbol{u_i}, \boldsymbol{\theta_b})^q}{q}, \qquad (5)$$

where y_b are personality traits predicted based on biased features, $p_y(\boldsymbol{u_i}, \boldsymbol{\theta_b})$ are probability assigned to the target traits

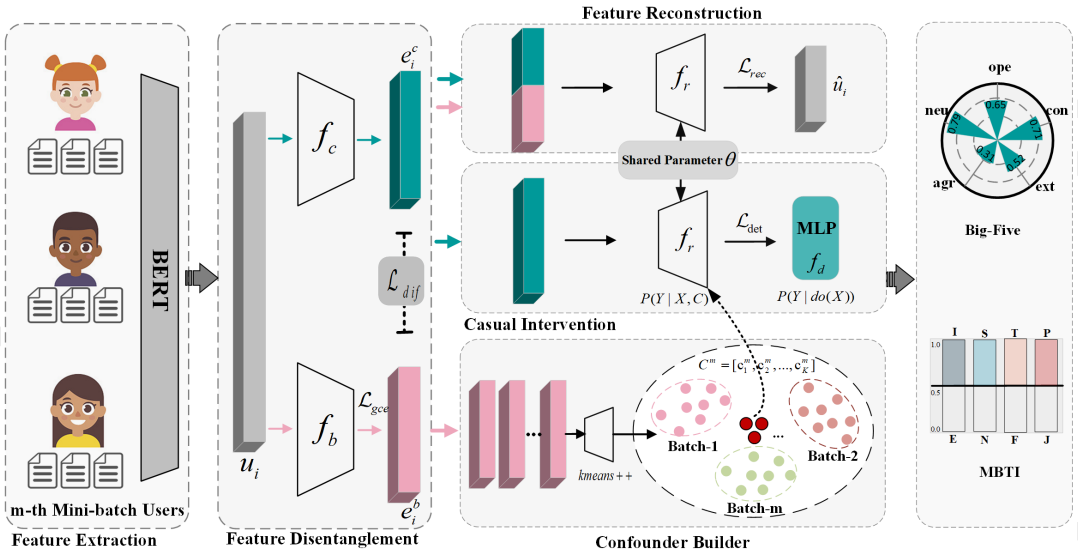


Figure 3: The architecture of our proposed IPDN model. In training, given m-th batch posts, the disentanglement and reconstruction module is used to decompose biased and causal features, and the decomposed biased features are aggregated into the confounder dictionary. In parallel, the reconstruction network are reused to approximate backdoor adjustment, ensuring that traits see each confounder prototype equally. In inference, given a test post, the confounders aggregated in the last iteration are combined and fed to f_c, f_r and f_d branches for detection.

y, and $q \in (0, 1]$ is a hyperparameter that controls the degree of amplification. The gradient of the GCE loss up-weights the gradient of standard Cross Entropy (CE) loss when the sample has a high prediction probability toward target trait:

$$\Delta\mathcal{L}_{gce}(\mathrm{Y}_b, \mathrm{Y}) = q\Delta\mathcal{L}_{ce}(\mathrm{Y}_b, \mathrm{Y}), \tag{6}$$

The GCE loss trains a biased projection function by giving more weight to "easier" samples with strong agreements between the predictions and the labels, which amplifies the "prejudice" of bias projection function. In this way we require no explicit labeling of confounders, but instead cheap and generalized confounders. While causal representations are enforced by a soft orthogonal constraint between output of causal and biased projection function:

$$\mathcal{L}_{dif} = \left\| \boldsymbol{e}_i^{c\mathrm{T}} \cdot \boldsymbol{e}_i^{b} \right\|_F^2, \tag{7}$$

where $\|\cdot\|_F^2$ is the squared frobenius norm. To ensure that the separated features are within a reasonable range, given the input of causal features $\boldsymbol{e}_i^{\boldsymbol{c}}$ and biased features $\boldsymbol{e}_i^{\boldsymbol{b}}$, the IPDN should be able to reconstruct the basic features $\boldsymbol{u}_i$:

$$\hat{\boldsymbol{u}}_i = f_r((\boldsymbol{e}_i^{\boldsymbol{c}}, \boldsymbol{e}_i^{\boldsymbol{b}}), \boldsymbol{\theta_r}), \tag{8}$$

The reconstruction loss is defined as mean squared error between $\boldsymbol{u}_i$ and $\hat{\boldsymbol{u}}_i$:

$$\mathcal{L}_{rec} = MSE(\boldsymbol{u_i}, \hat{\boldsymbol{u}}_i). \tag{9}$$

Confounder Builder

Confounder builder is designed to stratify confounder prototypes from separated bias features for backdoor adjustment. We cluster the learned biased features to approximate confounder prototypes $C = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_K]$, where K is the number of confounder, and each cluster $\boldsymbol{c}_k \in \mathbb{R}^d$ represents a confounder prototype. In practice, instead of updating clusters of the entire training set, we incrementally update clusters as the mini-batch iterates. In m-th iteration, the confounder dictionary $C^m = [\mathbf{c}_1^m, \mathbf{c}_2^m, ..., \mathbf{c}_K^m]$ are dynamic updates by K-Means++ $(\boldsymbol{e_i^b}), \boldsymbol{e_i^b} \in batch\{1, 2 \cdot \cdot, m\}$, where $\boldsymbol{e_i^b}$ is the biased feature from the previous m batch data. This incremental updating process allows for efficient and scalable confounder refinement during the training process.

Deconfounded Training with Intervention

As described in Section 3.3, the backdoor adjustment is to obtain a personality detection model $P(Y|do(X))$ that is unaffected by unobserved confounders C. Unlike mainstream NWGM, which uses linear models to approximate conditional probabilities, we perform backdoor adjustment during training, i.e., any personality trait is viewed equally to all confounding stratum by the following training objective:

$$\mathcal{L}_{det} = \mathcal{L}_{ce}(\frac{1}{K}\sum_{k=1}^{K} f_d(f_r(\boldsymbol{e_i^c}, \boldsymbol{c_k})), y), \tag{10}$$

where $\boldsymbol{c_k}$ is the confounder generated in each iteration of the training process, and given a decoupled causal feature $\boldsymbol{e_i^c}$, we reuse the reconstruction network f_r as encoder and combine each confounding feature $\boldsymbol{c_k}$ to forward K times. Each pattern of the unknown confounder is considered in the detection and the spurious correlation would not dominate. In general, our IPDN is end-to-end trained with a joint learning strategy, where a feature disentanglement task is to help learn confounders and an intervention personality detection task to pursue the causal effect collaboratively. The overall objective is to minimize both losses simultaneously,where λ is hyperparameters.

$$\mathcal{L} = \lambda\left(\mathcal{L}_{gce} + \mathcal{L}_{dif} + \mathcal{L}_{rec}\right) + \mathcal{L}_{det}, \tag{11}$$

Model Inference

Once the IPDN network is well-trained, given a test user, the personality traits are inferred through the f_c, f_r, and f_d branch.

$$\widehat{y} = softmax(\frac{1}{N}\sum_{k=1}^{N} f_d(f_r(f_c(\boldsymbol{u_i}), \boldsymbol{c_k^M})), y), \qquad (12)$$

where $\boldsymbol{c_k^M}$ is the confounder feature aggregated in the last iteration M of training phase. By equalizing various confounders, unbiased personality inferences are ensured to a certain extent.

4 Experimental Settings

4.1 Datasets

Kaggle dataset is collected from PersonalityCafe platform where people openly discuss their MBTI in daily communication. The Kaggle dataset comprises 8675 anonymous users, with each user contributing between 45 to 50 posts. Pandora is consist of Reddit posts, which labeled with MBTI and demographics (age, gender, and location) for 9067 anonymous users. Essays is a well-known stream-of-consciousness dataset [Pennebaker and King, 1999] consisting of 2468 anonymous volunteers with essays recorded their minds within a short time frame. Each user is marked with binary labels of Big-Five. These annotations are determined through a standardized self-report questionnaire. Following previous works [Yang *et al.*, 2021b; Yang *et al.*, 2023b], these datasets are randomly divided into 6:2:2 for training, validation, and testing respectively.

4.2 Baselines

For a comprehensive evaluations, we compare IPDN with the following 4 group baselines:**1) Post semantics models: LIWC** [Tighe *et al.*, 2016] pioneerly extracts the LIWC psycholinguistic features and employs SVM for detection. **RCNN** [Xue *et al.*, 2018] is a hierarchical CNN structure, which combines textual semantics and LIWC psychological features. **BERT$_{\text{finetune}}$** [Mehta *et al.*, 2020; Ren *et al.*, 2021] fine-tuned BERT to achieve the optimal configuration for personality detection.**2) Post structure models: SN+Attn** [Lynn *et al.*, 2020] is a hierarchical network that utilizes GRU to learn user representations from word-level and sentence-level sequences. **TrigNet** [Yang *et al.*, 2021b] is a novel graph attention network that aggregates different posts from each user by inherent psychological structures. **D-DGCN** [Yang *et al.*, 2023b] is a dynamic graph convolutional network that automatically learns the structure between user-generated posts. **3) LLMs-based models:** the "gpt-3.5-turbo-0301" version of **ChatGPT** is applied for personalty detection, the temperature is set to 0, making the outputs deterministic for the identical inputs. **TAE** [Hu *et al.*, 2024] distills the LLM's knowledge to enhance the small model for personality detection. **4) Debiasing text classification: CORSAIR** [Qian *et al.*, 2021] is a counterfactual framework for debiasing text classification, It addresses dataset and keyword biases by generating two counterfactual documents during inference. In our implementation, we use demographic lexicon [Sap *et al.*, 2014] as keywords. **NWGM** [Chen *et al.*, 2023] approximates the do-calculus for desired interventions at the feature level. In practice, the confounder dictionary x is also pre-defined demographic lexicon.

4.3 Implementation Details

We implement our IPDN in Pytorch 935 1.11.0 and train it on three NVIDIA GeForce RTX 2080 GPUs. We utilized the Adam optimizer (Kingma et al., 2017) and searched for the learning rate among $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$. IPDN are trained for 80 and 120 epochs in single-dataset and cross-dataset experiments, respectively. Early stopping strategy is employed for training. To initialize the post embeddings, we used the pre-trained language model BERT with the bert-base-cased architecture. The output dimensions of the mapping function are set to 200, 200, and 300 for the Kaggle, Pandora, and Essays dataset. The dimensions of the confounder prototype are the same as the output dimension of the mapping function, which is to facilitate feature-level computation. The size K of confounder dictionary $C = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_K]$ (i.e., the number of clusters) are set to 64, 128, and 64 for the three datasets, respectively. We search for the trade-off parameter λ are searched in $(0, 1)$ for different datasets.

5 Experimental Analysis

5.1 Main Results

The overall results are shown in Table 2 and the major findings can be summarized as follows:

First, we can observe that IPDN outperforms all baselines. Compared with the state-of-the-art method (TAE), IPDN achieves 1.81%, 1.08%, and 1.30% improvements in average Macro-F1 on Kaggle, Pandora, and Essays datasets respectively. The results verify the superiority of our model in personality detection. We attribute this success to two-fold factors: **1)** The users in these datasets have diverse individual attributes, and the generated posts come from uncontrolled real-world scenarios. In this situation, IPDN can effectively eliminate spurious correlations by causal intervention and get substantial gains. **2)** Unknown confounders in personality detection are well captured by batch-based progressive confounder learning during model training.

Second, we also conducted a comparison between IPDN and mainstream debiasing methods (CORSAIR, NWGM). Both CORSAIR and NWGM utilize the same BERT model as the backbone network and rely on predefined demographic dictionaries for counterfactual inference and backdoor adjustment. CORSAIR and NWGM outperformed BERT$_{\text{finetune}}$ in all datasets, indicating that explicitly modeling confounders in the network is effective. Furthermore, IPDN generally performs better than CORSAIR and NWGM. A reasonable explanation is that IPDN performs real intervention by treating confounder features as trainable parameters, while the predefined confounders in CORSAIR and NWGM may not be sufficiently representative.

Third, all baselines exhibit generally poor performance on the *Sensing* / *Ntuition* trait in both the Kaggle and Pandora datasets. This deviation arises from the imbalance in the distribution of this trait, leading to insufficient predictive capa-

Datasets	Traits	Methods										
		LIWC	RCNN	BERT	SN+Attn	TrigNet	D-DGCN	ChatGPT	TAE	CORSAIR	NWGM	IPDN
Kaggle	I / E	53.34	59.74	64.65	65.43	69.54	69.52	65.86	**70.90**	66.77	67.55	69.83
	S / N	47.75	64.08	57.12	62.15	67.17	67.19	51.69	66.21	69.89	70.79	**72.84**
	T / F	76.72	78.77	77.95	78.05	79.06	80.53	78.60	81.17	75.45	78.50	**81.31**
	P / J	63.03	66.44	65.25	63.92	67.69	68.16	63.93	70.20	69.98	68.09	**71.52**
	Average	60.21	67.25	66.24	67.39	70.86	71.35	66.89	72.07	70.52	71.23	**73.88**
Pandora	I / E	44.74	48.55	56.60	56.98	56.69	61.55	55.52	**62.57**	61.55	62.25	62.44
	S / N	46.92	56.19	48.71	54.78	55.57	55.46	49.79	61.01	59.11	57.07	**60.74**
	T / F	65.37	64.39	64.70	60.95	66.38	71.07	71.25	69.28	**73.64**	72.00	72.80
	P / J	56.32	57.26	56.07	54.81	57.27	59.96	60.51	59.34	58.86	60.22	**60.52**
	Average	53.34	56.60	56.52	56.88	58.98	62.01	59.27	63.05	63.29	62.89	**64.13**
Essays	AGR	47.50	46.16	54.72	56.97	57.11	57.36	55.48	58.72	59.05	60.16	**60.88**
	CON	52.00	52.11	56.41	55.47	54.70	57.20	57.78	57.25	**57.80**	56.06	57.02
	EXT	49.20	39.40	58.42	55.33	57.09	59.34	54.83	59.69	59.53	60.39	**60.57**
	NEU	50.90	58.14	56.36	58.26	59.15	59.06	57.44	60.13	60.11	60.90	**61.32**
	OPN	52.40	59.80	59.76	60.77	60.62	61.80	61.16	61.04	61.45	62.74	**63.56**
	Average	50.40	51.12	57.13	57.36	57.73	58.95	57.34	59.37	59.58	60.05	**60.67**

Table 2: Overall results of our IPDN and baseline models in Macro-F1 (%) score. Numbers in **bold** mean that the improvement to the best performing baseline is statistically significant (t-test with p-value < 0.05).

Methods	Kaggle		Pandora		Essays	
	Macro-F1	Δ	Macro-F1	Δ	Macro-F1	Δ
IPDN	**73.88**	-	**64.13**	-	**60.67**	-
w/o CI	68.12	5.76 ↓	59.32	4.81 ↓	58.78	1.89 ↓
w/o CI+FD	66.02	7.86 ↓	55.60	8.53 ↓	56.53	5.65 ↓
r/w BF	65.58	8.30 ↓	54.89	9.24 ↓	54.12	6.55 ↓

Table 3: Ablation study results on all three datasets, where "Δ" indicates the corresponding performance change, and "w/o", "r/w" mean removing and replacing a component from the original IPDN, respectively.

Training	Test	Methods	I / E	S / N	T / F	P / J	Ave
Kaggle	**Pandora 20%**	D-DGCN	53.37	49.25	61.26	56.88	55.19
		IPDN	**59.54**	**57.40**	**64.35**	**58.07**	**59.84**
Pandora	**Kaggle 20%**	D-DGCN	44.09	48.32	58.84	55.60	51.71
		IPDN	**51.25**	**55.22**	**60.84**	**56.21**	**55.88**

Table 4: Results of cross-dataset validation on the Kaggle and Pandora datasets.

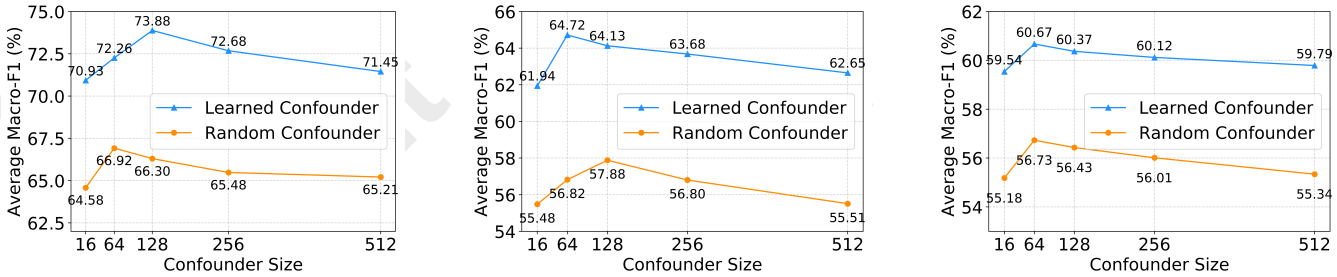
bility for the fewer traits. Fortunately, our IPDN strives to find contextual stratification confounders and refines the causal representation, leading to some improvements in this dimension. Moreover, it is worth noting that although causal intervention brings gains for all traits across three datasets, the improvements on the Essays dataset are not as pronounced as Kaggle and Pandora. We conjecture that this could be due to a limited training sample, resulting in insignificant confounding effects and insufficient intervention.

5.2 Ablation Study

We conduct thorough ablation studies for our IPDN on three datasets to verify the effectiveness of different components. As shown in Table 3, the performance drops by around 5.76%, 4.81% and 1.89% on Kaggle, Pandora, and Essays dataset after removing the causal intervention component (**w/o CL**) and feds the causality features $\boldsymbol{e_i^c}$ to the detection layer $f_d(f_r(\boldsymbol{e_i^c})$, which demonstrates the necessity of deconfounded training with intervention. Then, the performance continues to decline when the disentanglement component is further removed (**w/o CL+FD**) and uses the initial features u_i for detection $f_d(f_r(\boldsymbol{u_i}))$, highlighting the causal representations obtained by disentanglement are advantageous for personality detection. Conversely, directly employing biased features $f_d(f_r(\boldsymbol{e_i^b})$ yields the lowest performance (**r/w BF**), further emphasizing that our designed disentanglement module effectively isolates some hidden confounders.

5.3 Validation of Out-of-Distribution Scenario

We devise a challenging yet practical Out-of-Distribution (OOD) test configuration through cross-dataset validation. This is makes sense as the posts provided to recognition systems in real-world scenarios are produced by diverse individuals with varying preferences and in different environmental contexts. Consequently, the distribution of such data is beyond our control. Specifically, we use Kaggle as the training set and allocate 20% of Pandora as the test set, and vice versa. As shown in Table 4, IPDN shows an overall improvement compared to the strong baseline D-DGCN in cross dataset experiment. Although D-DGCN learns a rich document representation through post structure, there are more implicit confounders in cross-dataset testing. In contrast, IPDN isolates and removes these confounders through causal intervention, focusing more on causal relationships related to the target personality.



(a) Kaggle (b) Pandora (c) Essays

Figure 4: Visualization of detection performance as confounder size K increases.

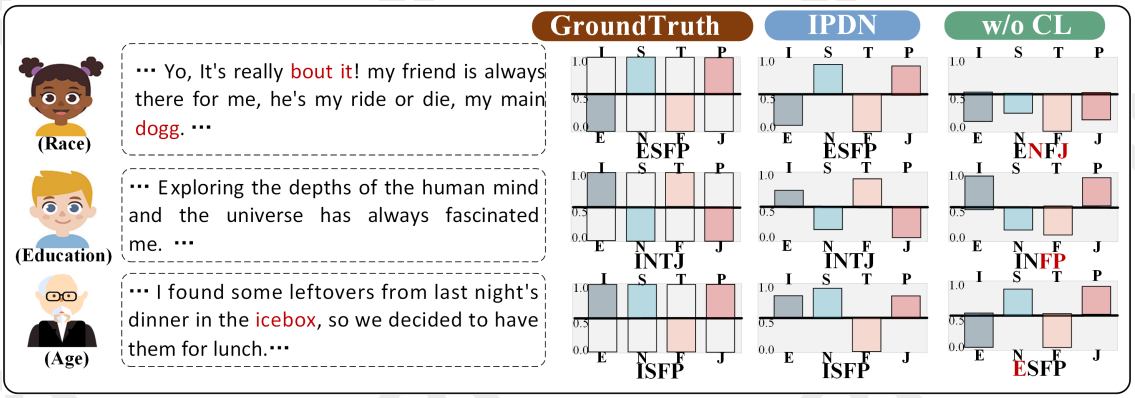


Figure 5: Differences between $P(Y|X)$ and $P(Y|do(X))$.

5.4 Effect of Confounder Prototype

To demonstrate the rationality of the learned confounder prototype, the random confounder dictionary is used for comparison. Simultaneously, to find the optimal size K for the confounder C, we conduct experiments on all datasets by setting K to 16, 32, 64, 128, 256, and 512, respectively. The results in Figure 4 indicate that the random prototype significantly hurts performance, thus confirming the effectiveness of our learnable confounder prototype. For datasets containing varying degrees of harmful bias, selecting an appropriate size K can effectively assist in deconfounded training.

5.5 Case Study of Causal Intervention

In Figure 5, we select three representative test cases to show the performance of the model before and after the intervention. Those cases are collected wildly from the MBTI community on Reddit, where people are mostly MBTI-tested with personal information. In the first row, this is a daily post shared by an African-American vernacular girl, and the popular slang ("dogg") has led the detection model to form a stereotype. In the second row, highly educated individuals typically have clear and logical language expression, which also serves as a shortcut in prediction. Furthermore, the main users of online social networks are young people, whereas, for older people, IPDN has untangled the false correlation between personality traits and some old-fashioned vocabulary ("icebox"), thereby improving the performance of the model.

5.6 Ethics Statement

The current research requires a thoughtful examination of ethical issues. 1) Privacy protection: This study is built on anonymized public datasets and was ethically vetted to comply with ethical guidelines and norms. 2) Risk of misinformation and misuse: Detection systems may be subject to misuse, e.g., discrimination, fraud, or manipulation. It is important to take steps to ensure that the detection results are used correctly, e.g., employment selection, loan approval, or legal judgments. and to prevent them from being misused or misrepresented. 3) Psychological impacts and individual rights: When personality assessment results in conflict with an individual's self-identity or social identity, it may harm self-esteem, personal image, and social relationships. Attention should be paid to mental well-being and personal rights when employing these systems.

6 Conclusion

In this paper, we analyze and identify the individual bias in the user-generated posts for personality detection from the causality perspective. We propose an interventional personality detection network that approximate backdoor adjustment to eliminates the hidden bias. Our IPDN does not require predefined confounders and instead learns confounders from posts. Experiments on three real-world datasets verify that IPDN can effectively eliminate biases and improve personality detection. In future work, we are interested in addressing the personality trait label bias with causal inference theory.

Acknowledgments

This work was sponsored by Beijing Nova Program (20230484409), National Natural Science Foundation of China (62272322, 62272323,62206148), and Beijing Postdoctoral Research Foundation (2025-135).

References

[Chawla *et al.*, 2023] Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092, December 2023.

[Chen *et al.*, 2023] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 627–638, 2023.

[Chien *et al.*, 2022] Shih-Yi Chien, Chih-Ling Chen, and Yao-Cheng Chan. The influence of personality traits in human-humanoid robot interaction. *Proceedings of the Association for Information Science and Technology*, 59(1):415–419, 2022.

[Coltheart, 1981] Max Coltheart. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505, 1981.

[Digman, 1990] J M Digman. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1):417–440, 1990.

[Fang *et al.*, 2023] Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. On text-based personality computing: Challenges and future directions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10861–10879, 2023.

[Feder *et al.*, 2022] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

[Gjurković *et al.*, 2021] Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. Pandora talks: Personality and demographics on reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, 2021.

[Hu *et al.*, 2024] Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18234–18242, 2024.

[Kerz *et al.*, 2022] Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. SPADE: A big five-mturk dataset of argumentative speech enriched with socio-demographics for personality detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6405–6419, June 2022.

[Lynn *et al.*, 2020] Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316, 2020.

[Mehta *et al.*, 2020] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189, 2020.

[Myers, 1997] Isabel Briggs Myers. *Introduction to Type: A description of the theory and applications of the Myers-Briggs type indicator*. Vision Australia Student Support, 1997.

[Nam *et al.*, 2020] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[Pearl, 2009] Judea Pearl. *Causality*. 2009.

[Pennebaker and King, 1999] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[Qian *et al.*, 2021] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5434–5445, 2021.

[Ren *et al.*, 2021] Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532, 2021.

[Sap *et al.*, 2014] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1146–1151, 2014.

[Schafer and Project, 1991] D Paul Schafer and World Culture Project. *The cultural personality*. World Culture Project Markham, 1991.

[Tandera *et al.*, 2017] Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611, 2017.

[Tausczik and Pennebaker, 2010] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[Tighe *et al.*, 2016] Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28, 2016.

[Wei *et al.*, 2022] Yinwei Wei, Xiang Wang, Liqiang Nie, Shaoyu Li, Dingxian Wang, and Tat-Seng Chua. Causal inference for knowledge graph based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[Xue *et al.*, 2018] Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48:4232–4246, 2018.

[Yang *et al.*, 2021a] Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. Multi-document transformer for personality detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14221–14229, 2021.

[Yang *et al.*, 2021b] Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. Psycholinguistic tripartite graph network for personality detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4229–4239, 2021.

[Yang *et al.*, 2022] Qi Yang, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. Personality-driven social multimedia content recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7290–7299, 2022.

[Yang *et al.*, 2023a] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context deconfounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2023.

[Yang *et al.*, 2023b] Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904, 2023.

[Yang *et al.*, 2023c] Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, 2023.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Zhou *et al.*, 2023] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4227–4241, July 2023.

[Zhu *et al.*, 2022] Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. Contrastive graph transformer network for personality detection. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4559–4565, 7 2022.