小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0206
2026 年 5 月 第 5 期
Vol.47 No.5 2026

一种融合因果干预的新闻反事实去偏方法

易锦成,蒋少华,文启鹏

(湖南师范大学 信息科学与工程学院,长沙 410081)
E-mail:jiangshaohua@hunnu.edu.cn

摘 要: 鉴于新闻偏见对公众认知、社会信任及公平性的深层影响,利用自然语言处理技术构建透明、可解释的去偏见框架,已成为传播学与人工智能的交叉研究热点.现有研究主要围绕两类偏见展开:词汇偏见和框架偏见.在词汇偏见方面,主流方法多通过词汇替换来消除文本中的显性偏见词,但仍存在中性词语中蕴含隐性立场倾向、上下文适应性差等问题;在框架偏见方面,现有研究多采用文本重构或多文本融合生成的方式来建模中立文本,但存在框架偏见不可观测、立场冲突难解耦、生成目标模糊等挑战,限制了偏见缓解效果的进一步提升.针对上述问题,本文提出一种融合因果干预与反事实推理的多阶段新闻偏见缓解方法.首先,针对词汇偏见,构建基于 PMI 的多立场偏见词典,并引入后门干预机制,通过语义相似度匹配进行词语替换,从而缓解显性偏见.其次,为应对结构性框架偏见的不可观测性,本文引入反事实推理方法,基于因果公式 TIE = TE-NDE 建模偏左与偏右框架对中立表达的影响,其中 TE 表示总偏见效应,NDE 表示中立文本的自然直接效应,TIE 则反映偏见传播的间接效应.最后,本文引入一个预训练的偏见检测器作为辅助监督模块,增强生成模型对文本中立性与专业性的建模能力.实验结果表明,本文方法在多个偏见缓解与文本质量评估指标上均显著优于现有主流方法,验证了该方法在多源新闻文本去偏任务中的有效性与实用价值.

关 键 词: 新闻偏见;词汇偏见;框架偏见;因果干预;反事实推理
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1147-09

News Counterfactual De-bias Method Integrating Causal Intervention

YI Jincheng, JIANG Shaohua, WEN Qipeng
(College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China)

Abstract: Given the profound impact of news bias on public perception, social trust, and fairness, building transparent and interpretable debiasing frameworks with natural language processing (NLP) has become a research focus at the intersection of communication studies and artificial intelligence. Existing work mainly targets lexical bias and framing bias. For lexical bias, mainstream approaches replace explicit biased words but struggle with implicit stance tendencies in neutral words and poor contextual adaptability. For framing bias, text reconstruction or multi-text fusion is often used, yet faces the unobservability of framing bias, difficulty in disentangling stance conflicts, and vague generation objectives, limiting further improvement. To address these issues, we propose a multi-stage news bias mitigation method combining causal intervention and counterfactual reasoning. A PMI-based multi-stance lexicon and a back-door intervention mechanism perform semantic similarity-based word replacement to reduce explicit bias. Counterfactual reasoning with TIE = TE-NDE models the influence of left-and right-leaning frames on neutral expressions, where TE is the total bias effect, NDE is the natural direct effect, and TIE captures indirect bias propagation. A pre-trained bias detector provides auxiliary supervision, enhancing the model's ability to generate neutral and professional text. Experiments show our approach significantly outperforms mainstream methods across multiple debiasing and text quality metrics, confirming its effectiveness in multi-source news debiasing tasks.

Keywords: news bias; lexical bias; framing bias; causal inference; counterfactual reasoning

0 引 言

在信息爆炸的时代,新闻媒体在塑造公众认知、引导社会舆论方面扮演着至关重要的角色.然而,随着媒体立场的日益分化,新闻报道中的偏见问题日益严重.尤其在涉及敏感议题(如移民、气候、种族等)时,不同立场的媒体通过选词、措辞、叙事结构等方式影响读者判断,从而形成“回音室效应”[1]和“信息茧房”[2],这种现象被称为新闻偏见,它不仅破坏了新闻应有的客观性和公正性,也可能在舆论层面加剧社会撕裂、误导政策方向,甚至威胁社会制度的健康运行.

正因如此,如何识别并消除新闻文本中的偏见因素,构建中立、客观的文本或报道形式,成为当前自然语言处理领域的研究热点[3].在本文中新闻偏见缓解任务被定义为融合同一主题下包含不同偏见的新闻,生成该主题不带有偏见的新闻.其任务如图 1 所示.Lee[3]在 2022 年提出新闻偏见以词偏见和框架偏见两个途径单独或融合出现.单独的词汇偏见通常

收稿日期:2025-05-19 收修改稿日期:2025-08-19 基金项目:国家自然科学基金项目(62472170)资助. 作者简介:**易锦成**,男,2000 年生,硕士研究生,CCF 学生会员,研究方向为自然语言处理、偏见缓解、偏见检测;**蒋少华**(通信作者),男,1976 年生,博士,副教授,研究方向为自然语言处理、生物信息学;**文启鹏**,男,2002 年生,硕士研究生,研究方向为自然语言处理.

表现为使用带有情感色彩或立场倾向的词汇,例如使用“非法移民”代替“无证移民”,或使用“激进”而非“坚定”等,这类偏见可通过词典或情感标注方法识别并缓解[4]. 而框架偏见则更为隐蔽,通常体现在文本中选择性地强调或省略某些事实,或通过结构化的叙述方式引导特定解读. 相较于词汇偏见,框架偏见更加复杂,往往无法仅通过表层词汇特征捕捉,因此对其建模与控制仍面临诸多挑战. 其任务如图 1 所示.

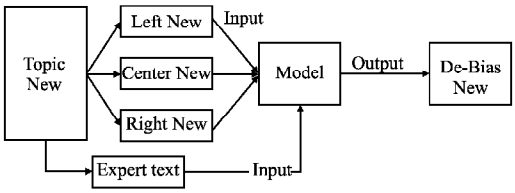


图 1 任务描述图

Fig. 1 Task description diagram

对于词汇偏见与框架偏见,本文在表 1 中展示本文所用的数据集案例以及其中包含的词汇偏见与框架偏见,以及不带有偏见的文本如何体现.

表 1 词汇偏见与框架偏见展示表

Table 1 Vocabulary bias and framework bias presentation table

类型	新闻
偏左新闻文本	The Trump administration's efforts to strip protections from more than half a million legal immigrants could devastate the health sector, endangering care for the elderly and worsening rates of both chronic and infectious diseases. Hundreds of thousands of health care workers, including an estimated 30,000 legal immigrants from Cuba, Haiti, Nicaragua and Venezuela, are at risk of being deported — worrying providers and patients who rely on them for everything from nursing and physical therapy to maintenance, janitorial, food-service and housekeeping work.
中立新闻文本	The immigration crackdown may already be starting to show up in the job market. Employment growth in industries that rely heavily on unauthorized workers has slowed. There has been a large decline in the foreign-born labor force since March. And recent immigrants appear more reluctant to take part in the Labor Department's monthly survey of households.
偏右新闻文本	President Trump signed so many executive orders on his first two days in office that many Americans were naturally left wondering how his policy changes would affect their everyday lives. One such question looms particularly large: If we deport illegal aliens, who will take their place in working low-skilled jobs? Many seem to think that the economy is completely dependent on illegal aliens working low-skilled jobs. According to the Brookings Institution, for example, deportations would lead to job losses among illegal aliens and U. S. natives alike. The deportation of the former would…
无偏专家文本	President Donald Trump's immigration crackdown has created questions for economists, employers, and workers as to how the decrease in migrant workers may impact the job market and the economy.

在表 1 中展示了同一主题下偏左、中立、偏右的偏见新闻文本以及专家撰写的无偏见文本,在偏左新闻文本通过选择性的强调某些事实、后果来影响读者的情绪或立场,其中通过“devastate the health sector”,“endangering care”,“worsening diseases”强调负面后果,这是通过框架偏见来达到形成偏见的目的. 在中立新闻文本中存在轻微的框架偏见,即通过“may already be starting to show up in the job market”这一句,隐含地将“immigration crackdown”与第 2 段第 1 句就业市场放缓形成一定的因果暗示,并强调负面影响. 在偏右新闻文本中使用了“illegal aliens”这种带有偏见的词汇来形成偏见. 在表 1 中可以得到的信息是词汇偏见可以通过替换偏见词来解决,但框架偏见需要对文本进行修改来解决.

为了在词汇偏见与框架偏见融合出现的情况下缓解新闻偏见,本文提出一种结合因果干预与反事实推理的生成建模框架,旨在系统地建模并消除新闻文本中的多层次偏见. 本文的方法包括 3 个核心步骤:

首先,本文通过因果干预消除词汇偏见带来的影响,以构建偏见词典为工具,混杂因子偏见词进行因果干预,而偏见词典基于大规模语料[5]构建. 通过进一步比较词语的互信息值与上下文语义相似性,本文识别出输入文本中的潜在偏见词,并采用中立词进行替换,从而构成对混淆因子的后门干预,以达到消除混杂因子影响的目的. 该过程的本质是以偏见词作为混淆因子,通过词汇替换来控制偏见源对生成结果的影响,从而估计去偏情况下的真实输出,即完成因果干预去除词汇偏见. 正如图 3(b)所示.

其次,为了消除框架偏见带来的影响,本文进一步引入反事实推理以建模并剥离文本中的非显性偏见影响. 本文构造了反事实输入,将文本中的潜在偏见因素归零,获得文本在“无框架偏见”情境下的预测输出,建模文本框架偏见,并基于因果推理理论,提出将偏左、偏右文本中所携带的框架偏见建模为总间接效应(Total Indirect Effect,TIE),并进行如下定义:

定义 1. 总效应(Total Effect,TE)定义为去除词汇偏见后的偏左、中立、偏右文本与真实中立输入的输出差异,即其所携带的框架偏见;

定义 2. 自然直接效应(Natural Direct Effect,NDE)表示在控制左、右文本输入不变的情况下,中立文本自身的框架偏见影响;

定义 3. 总间接效应(TIE = TE - NDE)即为在控制词汇偏见与文本框架偏见后,建模偏左、偏右文本所携带的框架偏见.

通过上述建模方式,本文不仅有效地区分了词汇偏见与框架偏见,还实现了对多源输入(偏左、偏右、中立文本)中不同层次偏见的因果干预. 最终,本文将构造出的偏见无关特征输入解码器,并引导其生成更加中立、专业的专家级文本,从而推动新闻自动生成系统的公平性与可信度.

最后,为了进一步提升生成文本的中立性与专业性,本文还引入了一个预训练的偏见检测器作为辅助模块. 该检测器以大量标注数据进行训练,能够有效识别文本中隐含的政治立场倾向(如偏左、偏右或中立),不仅可用于模型训练阶段的偏见判别,还可作为额外的监督信号对生成结果进行评价与约束.

本文的主要贡献总结如下:

1)提出了一种基于因果干预的词偏见消除方法,并通过 PMI 算法从大量偏见文本中提取偏见词构建偏见词典;

2)提出结合反事实推理的框架偏见建模方法,通过因果效应分解公式 $TIE = TE - NDE$ 建模并量化偏左与偏右文本的框架偏见;

3)训练并加入一个偏见检测器,通过有监督的方式提升生成文本的中立性与专业性.

1　研究现状

1.1　因果推理

因果推断理论自 Pearl[6] 提出以来,已发展成为消除数据偏见的重要统计建模技术[7]. 该技术能够有效识别变量间的虚假相关和混杂效应,为数据去偏提供了科学依据. Glymour[6] 等人系统性地将因果干预和反事实推理纳入因果推断框架,极大地拓展了其应用范围.

在方法论层面,研究者们提出了两大主要技术路线:1)基于潜在结果框架的反事实结果预测方法[8];2)基于结构因果模型的 do-操作[7]. 特别是 Pearl 等人[9] 提出的后门调整技术,为实际应用中的数据去偏问题提供了可操作的解决方案.

这些方法已在多个领域取得显著成效. 在推荐系统领域,Zhang 等人[10] 和 Wang 等人[11] 成功应用因果干预消除用户行为数据中的选择偏差;在自然语言处理领域,Tian 等人[12] 在自然语言推理任务中,Qian 等人[13] 在文本分类任务中,Zhang 等人[14] 在命名实体识别任务中,都验证了因果方法的有效性;Li 等人[15] 则将因果推断应用于预训练语言模型的去偏研究. 而 Zhu 等人[16] 在虚假新闻检测中通过反事实推理成功消除了文本实体分布带来的偏差,本文重点运用因果干预与反事实推理建模文本偏见.

1.2　媒体偏见缓解

近年来,偏见缓解成为自然语言处理领域的研究重点之一[17]. 一部分工作多聚焦于某一特定类别的偏见问题,例如 Sun 等人[18] 对性别偏见的探讨,以及 Lei 等人[19] 对政治偏见的研究,但其用于检测任务,仅限于句子级别分析,难以捕捉跨句的隐性框架偏见. 顾亦然等人[20] 通过融合数据集进行风格迁移研究,但数据集融合可能引入噪声或风格冲突,对特定领域的适配性不足. 在方法上,Manzini 等人[21] 提出了识别与去除词嵌入中多类别偏见的技术,但无法处理上下文依赖的语义变化,而 Bordia 和 Bowman[22] 则致力于减轻词汇级语言模型中的性别偏见,但其领域与偏见类型单一.

随着研究的深入,学者们提出了多种用于文本偏见消除的策略. Pryzant 等人[23] 和 Madanagopal 与 Caverlee[24] 通过对偏见表达的局部修改(如替换词语或句子)以降低文本的偏见倾向,然而替换的词语或句子在文中会出现上下文适应性差等问题. Liu 等人[25] 则基于 Transformer 模型结构构建生成对抗网络框架,用于中和新闻中的政治立场极性,而“中立”的新闻并不表示没有偏见,仍然包含框架偏见. 王剑等人[26] 则通过词对齐的对抗学习进行跨语言文本生成,然而会出现特定的文本难以翻译的问题.

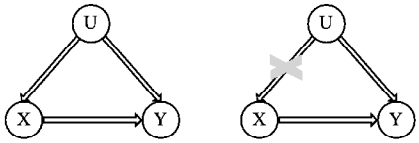
目前尚缺乏整合了偏见极性标注与专家级中立文本的大规模新闻数据集,也尚未建立统一的偏见缓解评估体系,但 Lee 等人[3] 提供了一个主题新闻数据集并将其用于新闻去偏任务,每个主题包含 3 个不同立场的新闻文本,但该方法并没有建模框架偏见,难以有针对性地进行偏见缓解. 本文在该数据集上运用因果干预和反事实推理建模文本偏见,并进行偏见缓解.

2　相关知识

在这一部分,本文将对因果图,因果干预和反事实推理 3 部分进行解释说明.

2.1　因果图

因果图(Causal Graph)是一种以图结构形式编码变量之间因果关系的建模方法,通常表现为有向无环图(Directed Acyclic Graph,DAG). 在因果图中,节点表示随机变量,有向边表示因果影响的方向,即源节点对目标节点存在直接的因果作用. 因果图不仅揭示了变量间的结构性依赖,还为识别混杂因素、推断干预效应以及执行反事实推理提供了理论依据. 通过因果图,可以系统地应用如后门准则、前门准则等方法,来确定在特定因果假设下的可辨识性问题.



(a) 未干涉因果图　　(b) 干涉后因果图

图 2　因果图示例

Fig. 2　Example of cause-and-effect diagram

例如,图 2 展示了一个典型的因果图,表明 X 是 Y 的因果因素,而 U 是混杂因素. 对于 X 和 Y,有前门路径 $X\rightarrow Y$,有后门路径 $X\leftarrow U\rightarrow Y$.

2.2　因果干预

因果干预(Causal Intervention)指的是通过主动操控某一变量的取值,来打破原有的自然依赖结构,从而识别因果效应的过程. 与仅基于观测数据进行推断不同,因果干预通过“do-运算”(do-calculus)显式地改变系统中的变量状态,屏蔽掉混杂路径的干扰,进而准确估计目标因果关系. 在因果图的框架下,干涉操作允许人为设定变量为外部控制的固定值,并考察系统随之产生的响应,从而揭示直接效应、间接效应及自然直接效应(NDE)等关键因果量.

以图 2(a)为例,假设干预变量 X,并且存在混杂因子 U,则干预后的因果分布 $P(Y|\mathrm{do}(X))$ 可以表示为:

$$P(Y|X)=\sum_{u}P(Y|X,u)P(u) \tag{1}$$

其中,u 表示混杂因子 U 的值. 通过因果干预,可以切断混杂因子对输入 X 的影响.

2.3　反事实推理与因果效应

反事实推理(Counterfactual Reasoning)是一种用于推测在不同假设条件下的结果的统计推断方法. 通过反事实推理,本文可以估计处理变量对响应变量的因果效应.

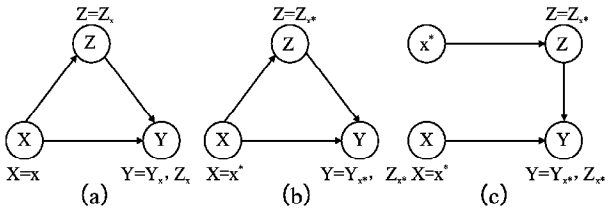


图 3　不同事实下的因果图

Fig. 3　Causal diagrams under different facts

反事实推理的基本思想是,通过对比两个相反的情境,来评估因果效应. 假设本文想要计算变量 X 对变量 Y 的总效应(Total Effect,TE). 在图 3 中,图 3(a)表示现实世界,计算 Y

的值时 $X=x$ 和 $Z=Z(X=x)$. 在图 3(b)中,X 被假设为 x^*,本文计算相应的反事实 Y. 总效应 TE 的计算公式为:

$$TE = Y_{x,Z_x} - Y_{x^*,Z_{x^*}} \tag{2}$$

其中,Y_{x,Z_x} 表示在 $X=x$ 和 $Z=Z(X=x)$ 的条件下计算 Y 的值,$Y_{x^*,Z_{x^*}}$ 表示在 $X=x^*$ 和 $Z=Z(X=x^*)$ 件下计算 Y 的值.

总效应 TE 可以分解为自然效应(NDE)和总间接效应(TIE). 以图 3(c)为例,自然直接效应 NDE 计算的是对的直接影响,当中间变量被阻断时,如图 3(c)所示,其具体公式表示为:

$$NDE = Y_{x,Z_{x^*}} - Y_{x^*,Z_{x^*}} \tag{3}$$

因此,总间接效应 TIE 可以表示为:

$$TIE = TE - NDE = Y_{x,Z_x} - Y_{x,Z_x} \tag{4}$$

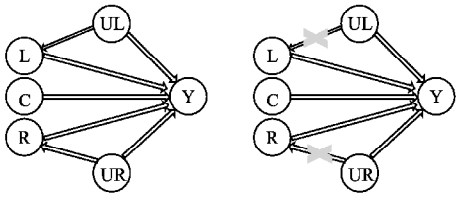
其中 TIE 表示 X 对 Y 的间接效应. 在本文中 TIE 被建模为输入文本的框架偏见.

3 方 法

在这一部分,本文首先将新闻偏见缓解任务描述为因果图,以清楚的反映因果之间的因果效应. 其次本文将根据 Lee[3] 在 2022 年提出单词偏见和语义偏见来分析如何根据因果干预和反事实推理来缓解新闻偏见.

3.1 新闻偏见缓解的因果图

新闻偏见缓解任务被定义为文本生成任务,即将一个主题下最具代表性的偏左、中立、偏右 3 篇新闻文本-标题拼接后作为模型输入,将该主题的名称、专家撰写的中立文本作为模型输出. 其中该主题名称被视为无偏标题,专家撰写的中立文本被视为无偏文本. 因此,新闻缓解任务的因果图如图 4(a)所示.



(a) 未干涉因果图 (b) 干涉后因果图

图 4 新闻偏见缓解任务因果图

Fig. 4 Causal diagram of the news bias mitigation task

其中 L、C、R 分别表示偏左、中立、偏右的新闻文本-标题,U_L,U_R 分别表示 L、R 的混杂因素,Y 表示 L、C、R 拼接输入 BART 后,其编码器输出的特征张量.

如图 4(a)所示,L 到 Y 存在前门路径 $L\rightarrow Y$,后门路径 $L\leftarrow U_L\rightarrow Y$,对于 C 到 Y 只存在前门路径 $C\rightarrow Y$,对于 R 到 Y,存在前门路径 $R\rightarrow Y$,后门路径 $R\leftarrow U_R\rightarrow Y$,因此本文对输入输出概况为 Y 受到输入 L,C,R 影响,其公式如公式(5)所示:

$$Y = Y(L,C,R) \tag{5}$$

其中 $Y(L,C,R)$ 可以进一步表示为:

$$Y(L,C,R) = Encoder(Cat(L,C,R)) \tag{6}$$

$Cat(\cdot)$ 表示拼接函数,本文中该函数皆在第二维度进行拼接. 在本文中 Y 被定义为输入文本中所包含的词、框架偏见,首先消除词偏见.

3.2 偏见词典构建

Chen(2023)[27] 指出,可以通过 do 运算对输入文本施加因果干预,参照公式(1)以削弱混杂变量对结果的干扰. 然而,在新闻偏见缓解任务中,由于同一主题下不同立场的新闻样本相对稀缺,而偏见词在稀缺样本下的先验频率缺乏鲁棒性,从而导致公式(1)变得难以适用. 为此,本文参考了 Ruan[4] 在 2023 年所提出的运用 PMI 算法从数量庞大的偏见文本中构建偏见词典,并对其进行改进,识别偏见文本中偏见词以达到对输入文本进行 do 运算的目的. 该方法在去偏比上达到了 17.16% 的效果. PMI 算法公式表示为:

$$PMI(w,b) = log\frac{P(w,b)}{P(w)\cdot P(b)} \tag{7}$$

其中,w 表示为候选词,b 表示为新闻偏见标签,$P(w,b)$ 表示为词 w 出现在立场为 b 的文本中的概率,$P(w)$ 表示词 w 在整个语料库中出现的频率,$P(b)$ 表示为偏见 b 的文本占比. 但在实际应用场景时 $P(w,b)$ 难以计算,因此本文通过计算在不同立场中的次数,来计算词的 PMI 值. 如公式(8)所示:

$$PMI(w,b) = log\frac{count(w,b)\cdot N}{count(w)\cdot count(b)} \tag{8}$$

其中,$count(w,b)$ 表示词 w 在偏见为 b 的文本中出现的次数,$count(w)$ 和 $count(b)$ 分别表示词 w 和偏见 b 在语料库中的出现次数,N 表示语料库中词的总数.

由于 Lee[3] 发布的数据集中数据量不够充足,在词的角度上缺乏概括性,本文以 Kiesel 等人[5] 于 2019 年提出的 736047 条带有偏见或中立的新闻文本为数据集构建词典,其具体分布如表 2 所示.

本文采用 PMI 算法构建了偏左词典、中立词典、偏右词典. 在选定词典候选词时,本文将词性控制在名词、动词、副词、形容词,该选取标准 Ruan[4] 已经证明可以取得好的效果. 词典具体描述如表 3 所示.

表 2 词典构建文本描述表

Table 2 Dictionary constructs a text description table

	偏左文本	中立文本	偏右文本	合计
文本数量	271500	224913	239634	736047

表 3 词典描述表

Table 3 Dictionary description table

	偏左词典	中立词典	偏右词典
词数量	9510	11791	5844
举 例	Polluters colonial coerce	challenge addition test	wapo correctness manipulates

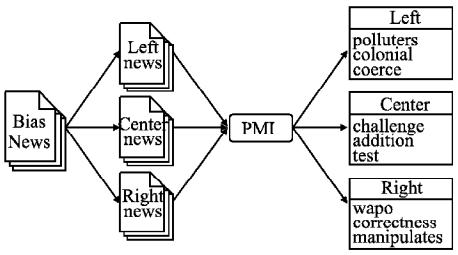


图 5 偏见词典构建示意图

Fig. 5 Schematic diagram of bias dictionary construction

根据表 2,可以发现偏左、中立、偏右文本在数量上并没

有表现出明显的差别,但是根据表 3 发现,词典所包含的偏左、偏右词数量具有显著差别,经统计:本文构建的包含词汇偏见与框架偏见的文本数量分布为 6:4. 由此,可以发现偏右文本比偏左文本更容易以框架偏见的形式隐含偏见. 词典构建流程如图 5 所示.

3.3　词替换方法

本文在 3.2 节中已构建偏见词典作为 do 运算的实行条件,本节通过偏见-中立词替换使混杂因子独立于输入文本,即 $U_L \perp L, U_R \perp R$,来达到阻断混杂因子影响的目的.

此外,为了解决偏见词在不同语境下可能存在情感或语义变化的问题,本文在偏见词典构建与词替换过程中引入了上下文敏感的词替换机制. 具体而言,本文不仅依赖 PMI 构建偏见词典捕捉具有统计性的立场偏见词,还引入了词语与上下文的语义相似度计算作为约束,确保替换词与原语义保持一致,从而减小替换对原文表达的破坏. 该算法主要分为 3 个步骤,现对其进行说明:

词特征编码:杨潇等人[28]曾通过矩阵编码进行同义词替换,而本文通过预训练的 Bart 模型(BART-large)[29]中的编码器将单词编码为特征张量,给定一个单词 w,其张量表示为:

$$V_w = BARTEncoder(w) \tag{9}$$

由公式(9)得偏左词典 $W_L = \{W_{L_1}, W_{L_2}, \cdots, W_{L_n}\}$,中立词典 $W_C = \{W_{C_1}, W_{C_2}, \cdots, W_{C_n}\}$ 偏右词典 $W_R = \{W_{R_1}, W_{R_2}, \cdots, W_{R_n}\}$ 的特征张量分别为:$V_L = \{V_{L_1}, V_{L_2}, \cdots, V_{L_n}\}$,$V_C = \{V_{C_1}, V_{C_2}, \cdots, V_{C_n}\}$ 以及 $V_R = \{V_{R_1}, V_{R_2}, \cdots, V_{R_n}\}$,同理可获得输入文本的特征质量其中 n 并不相同. 为了统一模型,本文全文基于预训练 BART-large 模型进行编码、解码.

相似候选词检索:以偏左的文本为例,输入一段偏左的文本 $S_L = \{w_1, w_2, \cdots, w_n\}$,对于每一个偏见词 $w_i \in W_L \cap S_L$,计算其与中立词典 W_C 中每一个词的余弦相似度:

$$sim_{word}(w_i, W_{C_j}) = \cos(V_{w_i}, V_{C_j}) \tag{10}$$

取相似度最高的前 K 个中立词作为候选词集合:

$$C(W_C) = \{W_{C_1}, W_{C_2}, \cdots, W_{C_K}\} \tag{11}$$

上下文保持性:为了确保替换后的文本在语义上尽可能接近原句,减小情感语义变化,本文进一步比较原始文本 S 与单词替换后的文本 S'的文本相似度与情感差异得分. 文本张量表示如下:

$$V_s = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{12}$$

其中 h_i 为 BART 编码器在每个 token 上的输出. 对于每一个候选词 w,将其分别替换原文中的偏见词并通过公式(10)计算句子相似度,除此之外通过集成的 BERT 模型作为情感分数计算工具,计算替换后的每一个句子与原文之间的情感差异得分,其公式表示如下:

$$p_i = Sent(S) \in [0,1] \tag{13}$$

其中 $Sent$ 表示为情感得分计算函数,p_i 表示句子的情感得分,替换前的句子与替换后的句子情感差异计算为 $d_e = |p_1 - p_2| \in [0,1]$,最后通过句子相似性得分与情感差异得分得到最终得分,表示如下:

$$S_{fused}(S,S') = \mathrm{sim}_{\mathrm{sent}}(S,S') \times (1 - d_e) \tag{14}$$

其中 S_{fused}表示最终的候选词分数,选取分数最高的词作为最终替换词. 偏右文本同理. 该方法能尽可能的保持替换后的上下文和情感与原文一致.

3.4　因果干预与去混杂因子训练

如图 3(a)所示,在 $L \to Y$、$R \to Y$ 分支中存在混杂因子 U_L,U_R(即偏左、偏右词),以 $L \to Y$ 为例,它通过学习可能性 $P(Y|L)$. 为了估计原始文本中偏见词对模型输出的因果效应,本文引入后门调整方法. 本文首先使用贝叶斯定理:

$$\begin{aligned} P(Y|L) &= \sum\nolimits_u P(Y|L,u_L)P(u_L|L) \\ &\propto \sum\nolimits_u P(Y|L,u_L)P(L|u)P(u_L) \end{aligned} \tag{15}$$

其中 $P(u_L)$表示偏见词的先验分布,$P(L|u_L)$表示偏见词 u_L 存在时生成文本 L 时的概率.

根据表 2 可知,偏见词是庞大的,而根据表 1 可知,数据集相对而言数量稀少,其并不能反映整个偏见文本的 $P(u_L)$,因此本文在 3.2 节中构建偏见词典并在 3.3 节中提出词替换算法,以此来切断 U_L 对 L、U_R 对 R 的影响,干预后 L'不再依赖 U_L,此时 $L \perp U$,即 U 与 L 解耦,因此 $P(u_L|L') = P(u_L)$,此时公式体现为:

$$\begin{aligned} P(Y|L') &= \sum\nolimits_u P(Y|L',u_L)P(u_L|L') \\ &\propto \sum\nolimits_u P(Y|L',u_L)P(u_L) \end{aligned} \tag{16}$$

其中 L'是不受混杂因子 U_L 影响的文本. 由于已经将偏左的词汇用词替换算法替换成了中立的词汇,替换后的文本 L'已经切断了 U_L 对 L 的因果影响,即 L'不再依赖 U_L. 此时 L'与 U_L 是独立的,正如图 4(b)所示,原始路径 $U_L \to L \to Y$ 已被切断,根据 d-分离准则,此时 Y 与 U_L 满足条件独立性,因此本文得到 Y 与混杂因素 U 条件独立,得到:

$$P(Y|L',u_L) = P(Y|L') \tag{17}$$

此时公式(16)可以进一步简化得到:

$$P(Y|L') = P(Y|L')\sum\nolimits_{u_L} P(u) = P(Y|L') \tag{18}$$

将公式(18)应用到偏右文本,可以得到 BART 编码器的输出:

$$Y = Y(Cat(L',C,R')) \tag{19}$$

为了防止模型学习到的是位置特征,在模型训练、测试时均将 L',C,R'随机顺序打乱,但为了方便表示,在本文中均以 L、C、R 的顺序出现. 因果干预后的因果图如图 3(b)所示.

本文将 TE 建模为偏左文本、中立文本、偏右文本所包含的总框架偏见,为此构建公式如下:

$$TE = Y_{L',C,R'} - Y_{C,C^*,C} \tag{20}$$

其中,C^*表示中立文本 C 的参考值,以标量 0 的形式表示.

该公式表示的是:在消除偏左文本 L 包含的词偏见、偏右文本 R 包含的词偏见的情况下形成 L',R',其与中立文本 C_s 输入模型的输出与理想状态下完全使用中立文本输入(即 $L' = C$、$R' = C$),且 C 被设为参考值 C^* 情况下输出之间的差异,从而衡量 L',C,R'三者所共同导致的总框架偏见效应.

3.5　通过反事实推理减轻中立文本偏见

截至目前,偏见词所带来的显性影响已在因果图中 $L \to Y$、$R \to Y$ 的路径上成功消除. 然而,基于图 6 所构建的新闻偏见缓解模型仍包含来自偏左文本、中立文本与偏右文本的框架偏见. 为进一步提高生成结果的客观性,本文引入反事实推理,以建模中立文本中隐含的框架偏差.

框架偏见通常表现为信息呈现方式的倾向性. 即便文本中未包含明显的偏见词汇,仍可能通过话语结构、语境构建、事件排序等隐性方式传递特定立场. 这类偏差具有不可观测性,难以通过显式特征加以建模. 因此,本文提出一种基于表示学习的反事实建模方法,从特征层面替换中立的表示,用以建模中立文本的框架偏见.

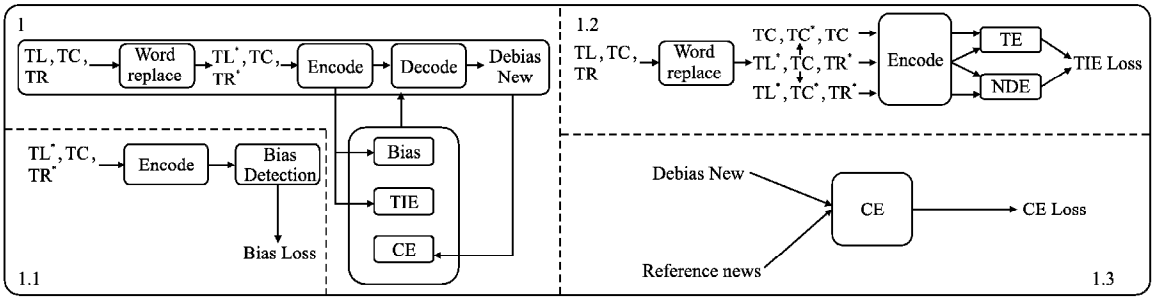


图 6 模型示意图

Fig. 6 Model schematic diagram

具体而言,对去除词偏见的偏左文本 L'、偏右文本 R' 以及中立文本 C,本文构造其对应的中立文本反事实表示,L' 和 R' 不变,而 C 替换为 C^*,从而建模中立文本的框架偏见.

中立文本框架偏见的影响可通过自然直接效应(NDE)行建模. 具体表示如下:

$$NDE = Y_{L',C,R'} - Y_{L',C^*,R'} \tag{21}$$

其中,C^* 表示将中立文本中的框架偏见剔除后的反事实表示,在本文中其被表示为 0 标量.

进一步的,可通过输入总框架偏见效应 TE 与中立文本框架偏见效应 NDE 得到偏左文本与偏右文本的框架偏见效应 TIE,其公式表示为:

$$TIE = TE - NDE = Y_{L',C,R'} - Y_{C,C^*,C} \tag{22}$$

3.6 模型训练

本文在图 4 说明了本文提出的因果干预去偏框架的训练与推理,在训练阶段本文根据 PoCa[30] 所提出的外部中立性约束方法,预训练了一个偏见检测模型,其在 Baly[31] 所提供数据上进行偏左、中立、偏右检测训练为三分类模型. 本文将其检测结果加入到损失函数中,因此模型损失函数为:

$$Loss = Loss_{FND} + \alpha Loss_{TIE} + \beta Loss_{bias} \tag{23}$$

其中 $Loss_{FND}$ 表示参考文本与生成文本之间的交叉熵损失函数,$Loss_{TIE}$ 表示输入文本中偏左、偏右文本的框架损失,$Loss_{bial}$ 表示偏见检测器结果与原标签之间的交叉熵损失.

4 实 验

为了评估所提出的方法,本文在 Lee[3] 所提出的数据集上进行了实验. 实验结果表明,模型在 5 个指标上都优于现有方法. 在接下来的小节中,本文以此介绍了数据集、损失函数、评价指标和实验配置,然后展示在 NeuS 数据集上的结果.

4.1 数据集

当前新闻偏见缓解领域所用的数据集相当稀少并且大部分数据来自 Allside. com 网站,其语言为英语,本文所用数据集来自于 Lee[3] 在 2022 年发布的数据集,该数据收集了来自 Allside. com 的 3066 组主题新闻,其中每组主题新闻包含内容为{主题名称,专家文本,偏左标题-新闻,中立标题-新闻,偏右标题-新闻}5 个元素,其中偏左新闻-标题、偏右新闻-标题是将 Allside. com 中的五类偏见归纳为 3 类偏见,即偏左、中立、偏右后的结果. 将数据集 NeuS 按照 8∶1∶1 的比例划分为训练集、验证集、测试集,如表 4 所示.

表 4 训练集表述表

Table 4 Training set description table

数据集	训练集(数量)		验证集(数量)		测试集(数量)	
	主题	新闻	主题	新闻	主题	新闻
NeuS	2452	7356	307	921	307	921

4.2 损失函数

损失函数用于衡量模型生成的文本与目标期望输出之间的差异,是模型优化和学习的基础. 在偏见缓解的任务中,损失函数有以下几个关键作用:

指导模型学习去偏特征(TIE):本文通过公式(22)定义了框架偏见的特征,损失函数利用 TIE 引导模型关注框架去偏特征,以生成在框架偏见上更加中立、客观的专家文本.

生成损失:本文采用生成文本与参考文本的交叉熵损失,用于衡量模型的生成文本与参考文本的差异,使生成文本更加接近专家撰写的去偏文本.

偏见识别损失:本文引入偏见识别损失作为辅助监督信号,通过预训练偏见识别模型增强模型的偏见敏感性与中立性生成能力,提高生成去偏文本时的准确性与鲁棒性.

具体而言,本文在公式(23)中已说明如何融合 3 个损失函数.

4.3 性能指标

当前新闻偏见缓解领域通常采用 RMSE、MAE、ROUGE1-Lsum 为评价指标,其中 RMSE 与 MAE 用于衡量生成文本与参考文本偏见标签的误差,ROUGE1-L 用于衡量生成文本质量,现对各个评价指标进行详细说明.

RMSE(Root Mean Squared Error)均方根误差在新闻偏见缓解领域指的是模型输出文本的偏见标签和参考文本(专家文本)之间的均方根误差,该指标由于平方项的存在,对偏差大的样本更加敏感,能够有效反映模型是否出现“重大偏差”,同样由于平方项的存在该指标不适用于大偏差的任务,但适用于新闻偏见缓和领域. 具体计算公式如公式(24)所示:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{24}$$

其中 $\hat{y}_i$ 表示生成文本的偏见标签,y_i 表示参考文本(专家文本)的偏见标签.

MAE(Mean Absolute Error)平均绝对误差在新闻偏见缓解领域的用法与 RMSE 相同,相比于 RMSE,其不容易受到离群值的极端影响,更能反映总体上的平均误差,适用于衡量模型整体表现的稳定性,具体计算如公式(25)所示:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{25}$$

其中 $\hat{y}$ 与 y_i 同理,表示为生成文本与参考文本的偏见标签.

ROUGE-1 用于衡量生成文本与参考文本在单词级别上的重合程度,能够反映模型是否捕捉到关键词和核心概念. 具体计算如公式(26)所示:

$$\begin{cases} ROUGE-1_{recall}=\dfrac{\sum_{w\in C\cap R}Count_{match}(w)}{\sum_{w\in R}Count(w)} \\ ROUGE-1_{precision}=\dfrac{\sum_{w\in C\cap R}Count_{match}(w)}{\sum_{w\in C}Count(w)} \\ ROUGE-1_{F1}=\dfrac{2\times precision\times recall}{precision+recall} \end{cases} \tag{26}$$

其中参考文本为 R,生成文本为 C,w 表示 C 与 R 中的单词.

ROUGE-2 用于衡量生成文本与参考文本在相邻两个词上的匹配,相较于 ROUGE-1 更能反映语言连贯性和短语质量,具体计算如公式(27)所示:

$$\begin{cases} ROUGE-2_{recall}=\dfrac{\sum_{bigramw\in C\cap R}Count_{match}(bigram)}{\sum_{bigram\in R}Count(bigram)} \\ ROUGE-2_{precision}=\dfrac{\sum_{bigramw\in C\cap R}Count_{match}(bigramw)}{\sum_{bigramw\in C}Count(bigramw)} \\ ROUGE-2_{F1}=\dfrac{2\times precision\times recall}{precision+recall} \end{cases} \tag{27}$$

其中生成文本为 C,参考文本为 R,以 bigram 为基本单元,也就是两个相邻词.

ROUGE-L 考虑了两个文本之间的最长公共子序列(LCS),能够评估结构性、语法完整性以及是否保留了原始表达顺序,具体计算如公式(28)所示:

$$\begin{cases} ROUGE-L_{recall}=\dfrac{LCS(C,R)}{|R|} \\ ROUGE-L_{precision}=\dfrac{LCS(C,R)}{|C|} \\ ROUGE-L_{F1}=\dfrac{2\times precision\times recall}{precision\times recall} \end{cases} \tag{28}$$

其中,LCS(C,R)表示生成文本与参考文本的最长公共子序列长度.

4.4　实验设置

本文使用 PyTorch 框架实现了模型,模型的输入文本为左、中、右 3 个文本、标题并用提示词"TITLE = >"和"ARTICLE = >"进行拼接,本文将输入的左、中、右 3 种文本顺序随机打乱,防止模型学到的是位置特征. 训练批次大小(Batch Size)为 16,测试批次大小为 4,学习率(lr)设置为 3e - 5. 模型均训练 3 个轮次(epoch),这些设置皆与 Lee[3] 设置一样.

硬件环境方面使用的是 NVIDIA A800 GPU,每次训练使用一块 GPU,显存为 80GB,训练框架为 PyTorch1.11.0.

4.5　实验结果

4.5.1　消融实验

为了评估模型中各个组成部分的贡献,本文进行了消融实验,通过系统性的逐步增加模块,评估其对模型生成性能的影响. 实验中分析的组件包括因果干预与词替换模块、反事实推理模块、偏见检测模块. 所有实验在相同的训练条件下进行,性能评估使用 MAE、RMSE、ROUGE1-L.

在消融实验中,均使用 NeuS 相同的数据设置办法对模型进行训练,并在验证集上进行评估. 本文采用预训练模型 BART-large 作为基准模型,其结果如表 5 所示.

表 5　消融实验表

Table 5　Ablation experiment table

模型	RMSE ↓	MAE ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
Baseline	0.2172	0.1666	39.09	18.93	29.73
+ CI	0.1935	0.1468	42.35	19.75	32.65
+ CI + CR	0.1825	0.1368	43.90	20.80	32.98
Ours	**0.1718**	**0.1298**	**44.50**	**21.30**	**33.75**

结果如表 5 所示. 完整模型在 RMSE、MAE、ROUGE-1、ROUGE-2、ROUGE-L 上提升了效果. 在增加指定模块后,模型性能出现了不同程度的增加. 其中加入 CI 表示加入以词替换的形式进行因果干预去除了混杂因子-偏见词对输出文本的影响,其在 RMSE、MAE 上分别带来了 0.0237、0.0198 的下降和 ROUGE-1、ROUGE-2、ROUGE-L 上分别带来了 3.26、0.82、2.92 的提升,加入 CR 表示通过框架偏见进行建模缓解文本框架偏见带来的影响,其在 RMSE、MAE 上分别带来了 0.0110、0.01 的下降和 ROUGE-1、ROUGE-2、ROUGE-L 上分别带来了 1.55、1.05、0.33 的提升,加入整体偏见损失表示在有监督的情况下用预训练模型 BERT 检测生成的偏见缓解文本所包含的偏见,其在 RMSE、MAE 上分别带来了 0.0107、0.007的下降和 ROUGE-1、ROUGE-2、ROUGE-L 上分别带来了 0.6、0.5、0.77 的提升. 证明模型可以有效缓解文本带来的偏见.

4.5.2　对比实验

由于新闻偏见缓解领域研究历史相对较短,先前建立的可用于新闻偏见缓解比较的方法较少,本文遵循 Lee 等人[3] 数据处理方法,本文对比了的基线模型包含了最近 5 年的一些模型,包括 OpenAI 的 ChatGPT(2022 ~ 2023)、GPT-4(2023),以及 2024 年最新发表的 PoCa 模型,并进行说明:

LexRank[32] 是一种基于图的无监督抽取图模型,它根据图中心性选择句子. 节点是句子,边用 tf-idf 加权.

Pegasus[33] 是一个抽象模型,它在 Multi-News[34] 数据集上微调 Pegasus 大模型.

NeuS[3] 开发了一种抽象文本生成方法,该方法学习以从标题到文章的分层顺序生成文本.

ChatGPT 是一种大型语言模型,可通过提示生成抽象文本. 本文使用 gpt-3.5-turbo 版本来获取文本.

GPT-4 是另一种自动生成抽象文本的大型语言模型. 本文在 gpt-4 版本创建文本.

PoCa[30]在2024年开发了一种通过极性校准、内容保留、语言自然度3个方法来实现去偏的强化学习模型,其在3个数据上预训练了3个不同的二分类模型实现.

表6显示了NeuS数据集上不同模型的结果,本文的模型在表格中最后一行并进行加粗,可以看到在5个指标上,本文提出的模型均取得了最好的效果.

表6 对比实验表

Table 6 Comparative experiment table

模型	RMSE↓	MAE↓	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
LexRank	0.2282	0.1838	38.68	15.94	25.66
Pegasus	0.2810	0.2344	37.33	16.02	25.54
NeuS	0.2172	0.1666	39.09	18.93	29.73
ChatGPT	0.2552	0.2076	42.01	16.24	26.12
GPT-4	0.2626	0.2133	42.35	16.48	26.30
PoCa	0.1834	0.1389	43.68	20.70	31.98
Ours	**0.1718**	**0.1298**	**44.50**	**21.30**	**33.75**

LexRank方法仅通过从原文中抽取重要句子生成摘要,缺乏针对词汇偏见与框架偏见的处理机制,因此在RMSE、MAE及ROUGE1-L等指标上表现不佳.Pegasus虽具备较强的生成能力,但未引入专门的偏见识别与消除机制,生成文本易延续原文立场,导致RMSE与MAE指标最为不理想.NeuS方法虽通过融合不同立场的文本生成无偏见摘要,在一定程度上提升了信息多样性,但未显式建模与消除框架偏见,效果仍受限制.ChatGPT和GPT-4作为通用大语言模型,具备较强的语言生成能力与多样化表达,但由于缺乏任务定向优化,其偏见缓解性能尚有提升空间.PoCa方法在内容、语义与去偏3个维度采用预训练独立模型改善内容保留度、语义流畅性与去偏程度的效果,但未处理词汇偏见和框架偏见,因此整体效果仍低于本文方法.由表6可见,本文方法通过词替换算法实现因果干预,有效消除显性词汇偏见;同时利用反事实推理建模框架偏见,缓解隐性框架偏见;最终结合预训练偏见检测器进一步进行中立性约束,在各项指标上均取得最优表现.

4.5.3 α和β值的影响

为了分析参数α和β值对模型的影响,本文将其设置为不同的值进行实验以找到最好的参数值,将α和β值设置为$\{0.1,0.3,0.5,0.7,0.9\}$中的值并进行调整,在训练过程中始终保持α和β值的和始终为1.经过实验训练发现,当$\alpha=0.7$且$\beta=0.3$时,可以在数据集上获得令人满意的偏见缓解结果.其结果如表7所示.

表7 参数讨论表

Table 7 Paramcter discussion tablc

α:β	RMSE↓	MAE↓	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
0.1:0.9	0.1760	0.1344	44.20	20.98	33.48
0.3:0.7	0.1754	0.1366	43.73	20.86	33.43
0.5:0.5	0.1736	0.1317	44.15	21.02	33.18
0.7:0.3	**0.1718**	**0.1298**	**44.50**	**21.30**	**33.75**
0.9:0.1	0.1772	0.1451	43.22	20.42	32.88

可以看到模型在框架偏见和整体偏见在0.7和0.3时在语义相似度和去偏程度上达到平衡,随着的比重增加,当达到0.7时模型达到最好的效果,当比重过于大时模型效果下降.

4.5.4 结果分析

在这一部分展示本文模型改写效果,在表8中本文展示了一组数据集,表中自上而下分别表示改写前的偏左、中立、偏右的文本与参考文本,改写之后的文本.其中偏左的文本带有主观性词汇“Superized Court”,并在正文中使用了“politically explosive”暗示强烈政治性倾向以显示其偏见.而中立文本指出问题的合法性是核心争议点,但正文内容偏题,未直接

表8 生成文本表

Table 8 Generate a text table

偏左新闻文本输入	TITLE => Supersized Court takes up Trump administration's plan to asks about citizenship in census ARTICLE => The supersized Court added a politically explosive case to its low-statement diversion Friday, agree to decide by the end of June whether the Trump administration can add a question.
中立新闻文本输入	TITLE => Supreme Court Agrees to Decide Legality of Census Citizenship Question ARTICLE => The 10-year questionnaire to U. S. residents is used to allocate government funding and reshape the number of congressional seats.
偏右新闻文本输入	TITLE => Supreme Court to decide whether citizenship question can be included in 2020 census ARTICLE => The Supreme Court will decide whether the 2020 census can include a question about citizenship, ensuring a quick review of a rival court ruling that blockcd thc Trump administration.
专家文本	TITLE => Supreme Court to Decide on Census Citizenship Question ARTICLE => The Supreme Court announced they will decide if the Trump administration can ask about citizenship status on the 2020 census.
生成文本	TITLE => Supreme Court to Decide Citizenship Question ARTICLE => The Supreme Court will decide whether the 2020 census can include a question about citizenship, ensuring a quick review of a lower court ruling that blocked the Trump administration from adding the question.

说明最高法院将快速处理此案.在偏右文本中语言上较为平和,但正文使用了“rival court ruling that blocked the Trump administration”,带入了具体政治动作,略显倾向.在本文模型生成文本中表达了主题,并且核心信息相同,与专家文本表达方式相近,语言风格较为中立.这说明本文提出的模型在新闻去偏缓解领域取得了令人满意的结果.表中的提示词与原数据集设置相同.

5 总 结

本文提出了一种基于因果干预与反事实推理的新闻偏见缓解模型,通过实验证明该模型能够在新闻缓解任务中取得优异性能.该模型通过偏见词典构建识别输入文本中的偏见词,并通过因果干预与词替换算法使得输入文本与混杂因子解耦,达到阻断混杂因子对输入文本影响的目的.除此之外,模型通过因果关系中的总效应与自然直接效应得到总间接效应,建模不可见的框架偏见,并将其作为损失函数加入模型使其在训练中降低.为了有监督的生成去偏文本,还在模型中加入在其他偏见新闻数据预训练的模型用于文本偏见检测.

但该模型同样存在局限性,无法获取消除框架偏见的中立文本使得建模框架偏见特征相对模糊,除此之外,数据集数量同样是当前工作的一大难题,在未来工作中本文将着力于数据集的扩充,以达到在完整新闻的情况下对新闻偏见进行缓解.

References:

[1] Cinelli M, De Francisci Morales G, Galezzi A, et al. The echo chamber effect on social media[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(9): e2023301118, doi: 10.1073/pnas.2023301118.

[2] Wang Y. An analysis of the information cocoon effect of news clients: today's headlines as an example[J]. Frontiers of Society, Science and Technology, 2023, 5(9): 7-11.

[3] Lee N, Bang Y, Yu T, et al. NeuS: neutral multi-news summarization for mitigating framing bias[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 3131-3148.

[4] Ruan Q, Namee B M, Dong R. Reducing media bias in news headlines[C]//Proceedings of the 31st Irish Conference on Artificial Intelligence and Cognitive Science, 2023: 1-4.

[5] Kiesel J, Mestre M, Shukla R, et al. SemEval-2019 task 4: hyperpartisan news detection[C]//Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis: Association for Computational Linguistics, 2019: 829-839.

[6] Pearl J. Causal inference in statistics: an overview[J]. Statistics Surveys, 2009: 396-146, doi: 10.1214/09-SS057.

[7] Wu A P, Kuang K, Xiong R X, et al. Instrumental variables in causal inference and machine learning: a survey[J]. ACM Computing Surveys, 2025, 57(11): 292, doi: 10.1145/3735969.

[8] Pearl J, Glymour M, Jewell N P. Causal inference in statistics: a primer[M]. Hoboken: Wiley, 2016.

[9] Blumberg, Joyce C. Causal inference for statistics, social, and biomedical sciences[J]. International Statistical Review, 2016, 84(1): 159, doi: 10.1111/insr.12170.

[10] Zhang Y, Feng F L, He X N, et al. Causal intervention for leveraging popularity bias in recommendation[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 11-20.

[11] Wang W, Feng F, He X, et al. Clicks can be cheating: counterfactual recommendation for mitigating clickbait issue[C]//44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, doi: 10.1145/3404835.3462962.

[12] Tian B, Cao Y, Zhang Y, et al. Debiasing NLU models via causal intervention and counterfactual reasoning[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022: 11376-11384.

[13] Qian C, Feng F, Wen L, et al. Counterfactual inference for text classification debiasing[C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, doi: 10.18653/v1/2021.acl-long.422.

[14] Zhang W, Lin H, Han X, et al. De-biasing distantly supervised named entity recognition via causal intervention[J]. arXiv preprint arXiv: 2106.09233, 2021.

[15] Li S B, Li X G, Shang L F, et al. How pre-trained language models capture factual knowledge? A causal-inspired analysis[C]//Findings of the Association for Computational Linguistics, 2022: 1720-1732.

[16] Zhu Y C, Sheng Q, Cao J, et al. Generalizing to the future: mitigating entity bias in fake news detection[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022: 2120-2125.

[17] Lei Y Y, Huang R H. Identifying conspiracy theories news based on event relation graph[C]//Findings of the Association for Computational Linguistics, 2023: 9811-9822.

[18] Sun T, Gaut A, Tang S, et al. Mitigating gender bias in natural language processing: literature review[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, doi: 10.18653/v1/P19-1159.

[19] Lei Y, Huang R, Wang L, et al. Sentence-level media bias analysis informed by discourse structures[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022: 10040-10050.

[20] GU Y R, XUE Y C, ZHANG T F. ID4TST: text style transfer model based on fused datasets[J]. Journal of Chinese Computer Systems, 2024, 45(10): 2338-2344.

[21] Manzini T, Lim Y C, Tsvetkov Y, et al. Black is to criminal as caucasian is to police: detecting and removing multiclass bias in word embeddings[J]. arXiv: 1904.04047v3, 2019.

[22] Bordia S, Bowman S R. Identifying and reducing gender bias in word-level language models[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 7-15.

[23] Pryzant R, Martinez R D, Dass N, et al. Automatically neutralizing subjective bias in text[C]//Association for the Advancement of Artificial Intelligence, 2020: 526-534.

[24] Madanagopal K, Caverlee J. Reinforced sequence training based subjective bias correction[C]//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, 2023: 1234-1245.

[25] Liu R, Jia C, Vosoughi S. A transformer-based framework for neutralizing and reversing the political polarity of news articles[C]//Proceedings of the ACM on Human-Computer Interaction, 2021: 1-26.

[26] WANG J, ZHANG Y, YU Z T, et al. A Chinese-vietnamese cross-language summary generation method using word alignment semi-supervised adversarial learning[J]. Journal of Chinese Computer Systems, 2022, 43(5): 992-997.

[27] Chen Z, Hu L, Li W, et al. Causal intervention and counterfactual reasoning for multi-modal fake news detection[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 1234-1245.

[28] YANG X, LI F, XIANG L Y. Synonym replacement steganography algorithm based on matrix coding[J]. Journal of Chinese Computer Systems, 2015, 36(6): 1296-1300.

[29] Lewis M, Liu Y, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.

[30] Lei Y, Song K, Cho S, et al. Polarity calibration for opinion summarization[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 5211-5224.

[31] Baly R, Martino G D S, Glass J, et al. We can detect your bias: predicting the political ideology of news articles[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 4982-4991.

[32] Radev D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Qiqihar Junior Teachers College, 2004, 22: 457-479, doi: 10.1613/jair.1523.

[33] Zhang J, Zhao Y, Saleh M, et al. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization[C]//International Conference on Machine Learning, 2020: 11328-11339.

[34] Fabbri A, Li I, She T, et al. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, doi: 10.18653/v1/P19-1102.

附中文参考文献:

[20] 顾亦然,薛宇辰,张腾飞. ID4TST:基于融合数据集的文本风格迁移模型[J]. 小型微型计算机系统,2024,45(10):2338-2344.

[26] 王　剑,张　莹,余正涛,等. 使用词对齐半监督对抗学习的汉越跨语言摘要生成方法[J]. 小型微型计算机系统,2022,43(5):992-997.

[28] 杨　潇,李　峰,向凌云. 基于矩阵编码的同义词替换隐写算法[J]. 小型微型计算机系统,2015,36(6):1296-1300.