NLP模型中的变形测试和公平性违规行为的认证缓和

# 马平川1,2, 王帅1\*, 刘进2 1香港科技大学 2北京电子科技大学

[pingchuan@ieee.org。](mailto:pingchuan@ieee.org) [shuaiw@cse.ust.hk。](mailto:shuaiw@cse.ust.hk) [liujin@besti.edu.cn](mailto:liujin@besti.edu.cn)

摘要

自然语言处理(NLP)模型已经越来越多地被用于敏感的应用主业，包括信用评分、保险和贷款评估。因此，知道NLP模型所做的决策没有对某些亚人群的不公平偏见是至关重要的。在本文中，我们提出了一个新的框架，采用变形测试，一个成熟的软件测试方案，来测试NLP模型，并找到引发公平性违规的不犯罪输入。此外，受最近在机器学习的认证鲁棒性方面的突破的启发，我们将NLP模型的公平性在一个实际的集合中表述为*(↪Ll\_1EB↩，k*)-公平性，并相应地平滑模型预测以减轻公平性违规行为。我们使用流行的(com- mercial)NLP模型展示了我们的技术，并成功地标记了可能导致公平性违规的歧视性输入。我们通过添加经认证的公平性保证，以适度的成本进一步增强了评估的模型。

# 介紹

最近自然语言处理（NLP）的成就一直依赖于人工智能技术的逐步发展。到目前为止，机器学习(ML)模型已经成为大多数现实世界NLP应用的基石，如文本分类、信息提取和问题回答(QA)。特别是，NLP系统已经被商业化，用于各种现实生活中的目的（犯罪预测、信用评分等），这对底层模型的精度和可靠性提出了强烈的要求。

尽管取得了这一引人注目的进展，然而，最近的研究表明，当决策因对敏感属性内的某些亚人群（例如，特定种族、基因、国籍或性取向的亚人群）存在偏见而不公平时，人工智能模型很可能会遇到严重的法律或道德问题[Kusner*等*，2017；Bolukbasi*等*，2016；Sheng*等*，2019]。事实上，最近的研究指出，黑人更可能是

∗通讯作者

以在完成情绪分类任务时被归类为具有负面情绪[Rhue，2018]。对于情感分析（SA）任务，预测也可以通过与身份群体相关的令牌来改变[Garg *et al.* , 2019]。迄今为止，ML模型的*公平性*已经以许多不需要的和潜在的苛刻方式被侵犯，引起了人们对其在犯罪预测、信用评分和招聘等领域的快速采用的深切关注[Barocas*等人*，2019]。

**原有评论**



......有很多演员太过努力去搞笑......。

...... 有很多**女演员**太过努力搞笑......。

...... 有很多**中国**演员太努力搞笑了.......。

**偏颇的评论**

图1：情感预测中的公平性违规。

我们通过用大型电影评论数据集训练一个用于SA的CNN模型来介绍一个激励性的例子。1这里，我们提出了三个样本与相应的标签和分数。很容易看出，当样本中的敏感属性（包括- ing性别和国籍）被改变或明确指定时，情感预测的结果将发生变化，并导致潜在的非预期偏差。

现有的研究已经提出了各种技术来排除NLP模型中的公平性违规行为。然而，这些工作存在各种缺点，例如依赖于重量级（不可扩展）的统计或符号分析工具，使用预先定义的模板生成输入，或仅处理结构化的数据表[Galhotra*等人*，2017；Udeshi*等人*，2018；Aggarwal*等人*，2019]。受软件工程中的rel- evant研究的启发，我们倡导将模型公平性分析制定为一个特定的软件测试任务的fo- cus，其中典型的NLP模型被视为一个"黑盒子"。特别是，我们建议采用一种成熟的软件测试方案，命名为*变形测试- ing*(MT)，以暴露导致NLP模型不公平预测的歧视性输入[Chen *et al.* , 1998]。为了测试软体，MT的工作原理是：1)扰动任意输入以产生一组突变，2)检查给定的元态关系(MR)是否仍与突变后的输入保持一致。每个MR，通常是一个不变量，描述了必要的属性。

1https://ai.stanford.edu/∼amaas/data/sentiment/。

目标软件。

虽然定义MR来断言模型的公平性是直接的（见第2节），但生成扰动的自然语言输入来测试模型的歧视性是具有挑战性的。对于in- stance，现有的作品用手编码的映射规则（例如，"他"→"她"）识别和替换某些标记[Kir-

itchenko和Mohammad，2018]。然而，手工编码规则

是很难在任意的感性属性内充分探索子群，因此容易遗漏某些歧视性的国家。定义这样的映射也需要大量的人工努力。在本研究中，我们利用先进的NLP方法，系统地自动收集任意敏感属性内的子群，并构建映射来扰动*非结构化的*自然语言句子。我们进一步排除了"不现实的"扰动w.r.t.标准语言流畅度指标[Ge*等，*2018]。因此，本工作中暴露出的缺陷主要是由于流畅和常见的句子，大概意味着现实生活中的问题。

我们进一步提出了通过缓解公平性违反来增强错误的NLP模型的技术。到目前为止，已有一些关于缓解NLP模型中公平性失误的工作，例如，通过数据增强和无偏词嵌入- ding [Zhao *et al.* , 2018; Park *et al.* , 2018; Sun *et al.]*尽管如此，这些工作需要访问和重新训练erro- neous模型。在这项研究中，我们倡导一种新的fo-。

MT的优点在于它能够通过变形关系(MRs)来保证软件的正确性，每个MR表示软件的必要的、通常是不变的属性。每个MR表示软件的必要的，通常是不变的属性。例如，为了测试sin(*x*)的实现，我们不需要知道每一个输入*x*的预期输出，而是可以断言当大量突变时，MR *sin*(*x*) = *sin*(*π* - *x*)是否总是成立。

x.适当定义的磁共振减轻了对测试的需要。

cle(ground truth)，从而使软件测试的灵活性大大增强。现在我们定义本研究中所涉及的NLP模型公平性，并据此制定MT程序。让M : X → *Rn*是一个模型，它把一个自然的lan-guage句子*x*作为输入，模型输出表示为argmax M(*x*)。假设一个扰动函数P扰动*x*

到集合*S中，*使得对于每一个*x*′∈*S*，存在一个敏感的

属性A和*xp* ƒ= *x*′ ′，其中*p，p*′∈A.换句话说，P通过在*x*中插入或改变一个令牌从*p*到*p*′来扰动*x*，假设*p*和*p*′都是A内的子种群，*x*和*x*′deem一对判别性*输入*，万一它们的模型预测不完全相同呢？因此，我们检查 fol-

*p*

降低MR，以确定违反公平的行为。

argmax M(*x*) = argmax M(*x*′)

而相应的MT工作流程可以表述为：。

算法1：模型公平性违规的MT。

cus关于减轻预测阶段的公平性失败。

因此，所提出的技术甚至可以应用于黑盒模型。受最近在cer- tified ML鲁棒性方面的突破的启发[Lecuyer*等人，2019*；Mu*等人*，2019；Cohen*等人*，2019]，我们用*(↪Ll\_1EB↩，k*)-公平性来形式化公平性符号(参见第3.1节)，我们提出了一个可以搭载在(黑盒)NLP模型上以平滑其输出的neu- tral阶段：平滑后的输出被认证为保留我们的*(↪Ll\_1EB↩，k*)-公平性。

贡献我们提出了MT-NLP，这是第一个用于识别和缓解NLP模型不公平预测的自动化框架。MT-NLP被设计为一个流水线，以1)采用ad- vanced NLP技术来系统地探索任意自然语言句子中识别的敏感属性中的子群体，2)对有关识别的敏感属性的句子进行扰动，并对NLP模型执行MT，以检测公平性违规行为，3)在认证的环境下将模型公平性形式化为*(↪Ll\_1EB↩，k*)公平性，并相应地减轻不公平预测。由于所提出的框架*并没有*利用NLP模型的任何实现细节，它在黑盒环境下对任何种类的ML模型(如DNN或SVM)都是稳健的，因此可以直接分析任何(远程)模式。我们广泛的评估划分了由行业巨头提供的事实上的商业NLP服务和常用的本地NLP模型的能力，共暴露了2874个判别输入（其中441个来自商业NLP服务）。我们以适度的成本进一步加强了经评估的（商业）NLP模型与认证担保的关系。

# 变质测试(MT)

变形测试（Metamorphic testing，MT）已被广泛用于自动检测软件缺陷[Chen *et al.*，1998]。强度

输入：模型M : X → *Rn*，输入*x*，敏感属性A，知识图*tt* := (*V，E*)

输出：判别输入集*Sdis*

1 *Sdis* ← ∅;

2 *S* ← P（*x，*A*，tt*）。

3 foreach *x*′∈*S* do

4if argmax M() xargmax M(*x*′)，则

*5Sdis*=*Sdis* ∪{*x*′}。

6端

7结束

8返回*Sdis*。

与现有的研究一致[Galhotra *et al.* , 2017; Udeshi *et al.* , 2018]，敏感属性A（例如，性别或种族）由用户选择，以使设计更具适应性。MT-NLP在pertur- bation阶段利用知识图*tt*（在2.1节中很快介绍）。现在我们阐述一下扰动函数P的设计。

## 句子扰动器P

如上所述，实现自动化的工作流来识别和突变非结构化自然语言句子中的敏感属性是一项挑战，现有的工作是利用手工编码的映射规则，因此，发现缺陷的能力受到目前已知映射的限制。现有的工作是利用手工编码的映射规则，因此，发现缺陷的能力受到目前已知映射的限制。在此基础上，我们力图在不使用任何预定义映射的情况下，对敏感属性进行系统性的探索和突变。

给定一个自然语言写的句子，我们的观察揭示了两个主要的突变机会。1）名词去注意敏感属性中的子群，2）广告词将敏感属性分配给（中性）名词。受这一观察的启发，P在概念上抓住了这两点

名词和形容词标记进行扰动。高级工作流程介绍如下。

算法2：句子扰动器P

输入：输入*x*，敏感属性A，知识图谱

*tt* := (*V, E*)

输出：扰动集*Sc*

1 SC ← ∅;

2 *T*←PartOfSpeech(*x*)．

3 foreach *t*∈*T and t is noun* do

4if IsPerson(*t, tt*) then

*5SC*←*SC*∪AnalogyMutate(*x，t，*A*，tt*)．

*6SC*←*SC*∪ActiveMutate(*x，t，*A*，tt*)。

7端

8结束

9 SC ←FluencyFilter(*SC*)。

10返回*Sc*。

所提出的扰动程序以一个自然语言句子*x*，一个用户选择的敏感属性A（如种族），还以一个知识图*tt*作为in- puts。知识图用标签化的边缘连接单词，实现对自然语言的系统化理解。

在这里，我们利用了一个流行的知识图，Concept- Net [Speer *et al.* , 2017]，它是专门为NLP任务设计的。 ConceptNet上的每个节点代表自然语言的一个词 或短语，Con- ceptNet上的标签边，如IsA、UsedFor和CapableOf，去注意连接节点之间的关系。为了实现IsPerson、AnalogyMutate和ActiveMutate（即将推出），我们广泛使用Concept- Net上的边缘IsA，它可以透彻地探索给定节点的"属性"。例如，通过查询"gay"的IsA边缘，我们知道"gay"是"lgbt"的一种类型，这是一个敏感的属性，很可能引发公平性违规。注意，IsA代表的是不对称关系。在本文的其余部分，我们用IsA来表示前向查询，而*IsArev*则反之。

Perturbator P首先执行标准的部分语音。

POS）分析，为每个词分配一个POS标签（如名词、动词和广告词）[Manning*等*，2014]。名词标记将被保留使用（第3行）。然后，我们检查一个名词是否

token指的是一个人（第4行；IsPerson将很快被解释）。给定句子和精确定位的标记，我们提出了两种策略来生成测试输入，通过改变名词标记与AnalogyMutate（行5）或其相关的形容词标记与ActiveMutate（行6）。此外，虽然句子的"自然度"对测试没有直接影响。

测试违反公平性。例如，把"笔"改成"铅笔"很可能会让人感到困惑。因此，我们分析的应用范围被微调到了最常见的环境--人的相关因素。直觉上，"模型公平性"与人的属性高度相关。

让ConceptNet上表示"人类"的节点为*h*，为了决定*t*是否代表人类，实现了IsPerson，通过IsA边来检查*h*和*t*之间的可到达性。例如，"恐怖分子"和"女警察"都可以视为与人类相关的名词，因为"恐怖分子"节点直接连接到*h*，而"女警察"也可以通过图上的两跳到达*h*。同时，对于每一个IsA，ConceptNet都会分配一个"权重"来表示这种知识关系的置信度得分。在查询可到达性时，IsPerson根据经验设计使用权重大于1.0的边。除了*h*，其他一些名词也可以放心地代表人。因此，我们总结出一个集合，*Ih*（见附录A）。

*t*视为与人有关的名词。

名词，只有当至少有一个*h*′∈*Ih*是*t*可以达到的*k*跳（*k*在实现中配置为两个）。为了加快搜索速度，我们还根据经验总结了一些非人的

tokens作为IsPerson的早期停止标准。

类比突变

将扰动目标限制为只与人类相关的标记，逐步提高扰动句子的质量。尽管如此，提出一个灵活的去符号来突变名词tokens仍然是一个挑战。一般来说，虽然有字符级的映射可以用来突变某些词，例如，将"服务员"等名词中的"er"改为"ress"，但这些规则的通用性有限，可能在很多角落的情况下找不到类似的标记（例如，"actor"

→ "演员"，其中"演员"不以"er"结尾）。)

在这一步，我们提出了一个灵活有效的类比方法。

突变方法AnalogyMutate，该方法的灵感来自于著名的"词类比"技术在操作词嵌入向量方面的应用[Ethayarajh *et al.* , 2019]。直观地讲，如果两对词之间的关系相似，则两对词嵌入向量之间的距离也相似。

因此，给定*d*是名词标记，*p*、*p*′是敏感属性内的子群，我们得到*d*′，*d*的类比，其公式如下。

*W* (*d*′) = *W* (*d*) + *W* (*p*′) - *W* (*p*)

其中*W*（-）表示给定词的嵌入向量。

雌性

我们通过保留"流利"的合成大小的句子来增强测试输入的真实性（第9行）。因此，触发公平性违规的合成句子大概可以im- ply现实生活中的苛刻程度。我们实现了[Ge *et al.* , 2018]中提出的标准句子流利度量。

现在我们详细说明IsPerson的动机和设计。鉴于自然语言标记的高度灵活的含义，我们的观察表明，扰动*ar-。*

*z*

雄性

女演员

### *y*

演员

### *x*

*bitrary名词*并不总能产生有用的句子，用于

图2：类比词的平行四边形结构。

算法3：类比突变

输入：输入*x*，令牌*t*，敏感属性A；知识图*tt* :=（*V，E*），词嵌入表*W* : *t* '→ *v。*

输出。模拟突变组*S*

1 *S* ← ∅;

2 *Sp* ←*tt*.*IsArev*(A)。

3 *pt* ← *argminpi*∈*Sp* ↪Lo\_1C1↩*W*（*pi*）-*W*（*t*）↪Lo\_1C1↩2．

4 *Sp* ← *Sp* ← *Sp*{*pt*}//设置减法。

5 foreach *pi*∈*Sp* do

*6tM* ← *W* -1(*W* (*pi*) + *W* (*t*) - *W* (*pt*))。

*7x*′ ←*x*.replace(*t，tM* );

*8S*← *S*∪{*x*′}。

9结束

10返回*S*。

受词类比技术的启发，上述公式提供了一种根据某些感性属性突变*d*的有效方法。例如，假设*p*和*p*′是"性别"属性中的子人群(例如，*p*为"男性"，*p*′为"女性")，图2提供了一个按mu-的样本映射。

将"演员"与它的类似名词"女演员"相提并论，它可以潜在地--------。

tially trigger discrimination on "gender".同样，给定*p*为"基督教"，*p*′为"伊斯兰教"，给定的翻译可以将"教堂"映射为"清真寺"。

AnalogyMutate在算法3中定义。给定用户选择的敏感属性，我们首先查询knowl- edge图，向后获取其类型节点（第2行）。例如，假设"性别"是所选的敏感属性A，知识图*tt*可以帮助收集到性别的"类型"节点，包括"男性"和"女性"。我们

然后迭代*Sp*中的每一个元素，并找出其词嵌入与*t*距离最短的*pt*；我们排除这个词（第4行），因为这大概表示A内的一个与*t*高度相关的子群体（如"男性"）（如"演员"）。第5-9行将上面介绍的类比突变与*W*（-）的反函数具体化，并迭代*Sp*内的剩余- ing标记*pi*（如"女性"）。

活性突变

算法4介绍了算法2中使用的ActiveMutate。为了避免潜在的冲突，如果*t*前面已经有广告词，我们就排除它。这一点可以通过部分语音分析轻松完成，在算法4中被省略。此外，为了防止诸如"女演员"这样的尴尬扰动，我们检查 *t*

是"中性"的w.r.t.属性A.为此，我们通过查询知识图(第2行)提取给定敏感属性A内的子种群。然后，我们对每个元素*pi*∈*Sp*进行迭代，并利用函数HasIsAEdge来检查*pi*在图上是否可以被*t*到达（行4）；*t*是"中性"的，w.r.t. *pi*

如果在两者之间没有找到边。 如果*t*是中性的，我们重述

*Sp*（第8行），并在前面插入一个敏感的形容词（第9行）。

# 认证缓解

最近在认证鲁棒性方面的进展为针对对抗性考试的可证明防御提供了一个严格的框架。然而，一般来说，扩展现有的

算法4：主动变异

输入：输入*x*，token *t*，敏感属性A；知识图*t* :=（*V，E*）。

输出。主动变异集*S*

1 *S* ← ∅;

2 *Sp* ←*tt*.*IsArev*(A)。

3 foreach *pi*∈*Sp* do

4if *tt*.hasIsAEdge(*t, pi*) then

5return ∅。

6端

7结束

8 foreach *pi*∈*Sp* do

*9tM* ← *pi* + *t* //用助词*pi*修饰*t*。

*10x*′ ←*x*.replace(*t，tM* );

11S ← S ∪{*x*′}。

12结束

13返回S。

方法到具有离散域的NLP任务，因为对抗性句子的魔力在实践中是不可测量的。尽管如此，给定一个任意句子*x*，由于我们已经构成了它的扰动集*S*，集*S*确实可以实现认证缓解（即将介绍）。另外，虽然相关的工作提出了采用像区间边界传播这样的统计方法来实现验证NLP模型对对抗性词替换的鲁棒性[Huang*等*，2019；Jia*等*，2019]，但我们注意到，这些工作不能直接en- hance黑盒NLP模型。

* 1. *(↪Ll\_1EB↩, k*)-公平性。

在本节中，我们制定了由预期输出稳定性边界得出的单个样本的公平性[Lecuyer *et al.* , 2019]。

定义3.1 一个函数*f* : X → [0*，b*)*N*和一个sam- ple x0是*(↪Ll\_1EB↩，k*)公平性w.r.t.敏感属性A，如果∃*S* ⊆ XA*，*|*S*| = *k*，∀*x，x*′∈{x0} ∪ *S*受制于

*f* (*x*) ≤ *e↪Ll\_1EB↩f* (*x*′)(1)

其中XA表示x0在上的所有可能突变的集合

A，|*S*|表示*S*的卡性，且*↪Ll\_1EB↩*≥0。

直观地讲，该定义要求至少有以下的产出

*k*偏向突变由公平性参数*↪Ll\_1EB↩限定*。当然，*k*越高表示对所述公平性参数的信心越高。在特殊情况下，当*↪Ll\_1EB↩*=0时，所有*k次*突变的输出都是相等的。

## 中性层设计

通过*(↪Ll\_1EB↩，k*)-公平性公式的启用，本节提出了一个中立层的设计，它可以搭载(黑盒)NLP模型，并减轻其公平性的违反。为此，我们重用测试阶段生成的扰动句子集*S*。

受随机响应的二进制版本[Warner，1965]以及如何通过楼梯机制[Kairouz*等*，2016]将其泛化为多值版本的启发，我们设计了一个认证的缓解机制，以实现*(↪Ll\_1EB↩，k*)公平性，具体如下。

输入样本。

类比Nutation。

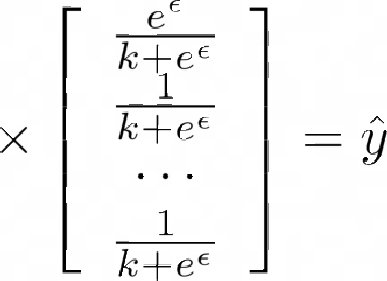
类比Nutation。

Nodel

积极的营养化。

输出层

60

40

频率(%)

平滑20

产量

直方图

累计

直方图

100

80

60

40

20

图3：认证缓解机制的工作流程。



定理1.*给定一个模型*M*、一个输入x及其相关的扰动集S，如果认证的模型f是无-的(↪Ll\_1EB↩，k)公允性保证，则*

0

0 0 25 50 75 100

扰动次数

图4：扰动分布。

*f* (*x*) =

eM(*x*) + Σ 1M(*x*′)(2)

*k* + *e↪Ll\_1EB↩*

*ǫ*

|  |  |  |
| --- | --- | --- |
| NLP模式和服务 | # 测试的数量  句子 | # 公平的数量  违规行为 |
| Amazon Comprehend API | 28,425 | 197 |
| 谷歌NLP API | 30,504 | 140 |
| 微软NLP API | 28,721 | 104 |
| 美国有线电视新闻网 | 174,183 | 1,111 |
| 逻辑回归(LR) | 174,183 | 1,322 |

*x*′∈*S*

*k* + *e↪Ll\_1EB↩*

认证机制为{*x*} ∪*S*中的每个元素分配权重，并产生一个平滑的输出*y*ˆ。 总的来说，这种方法不需要改变模型训练过程，但可以在推理时实现可证明的缓解。

box）模型。图3说明了在输入样本x0上的缓解过程，使用其相应的扰动参量。我们在附录B中进一步给出了关于其公平性保证的定理的证明。1关于其公平性保证的证明，见附录B。

如图3所示，所提出的缓解措施导致了额外的计算开销。总的来说，对于每个输入*x*，我们计算它的扰动集*S*，会产生*k个*额外的预判（recall |*S*| = *k*）。然而，由于每个样本的预测阶段是独立的，因此将*k*+1个预测阶段paral- lelize并减少处理量是可行的。

时候。

# 评价

表1:结果概述

图4研究了扰动结果，报告了从一个输入句子*x*中得出的扰动数的分布情况。我们报告说，大约67.0%的输入有5个以上的扰动，37.2%的输入有10个以上的扰动。一个输入的最大perburbation集是116个。总的来说，这些结果表明，所提出的技术可以从任意句子中自动合成大量的per- turbations，从而使系统----------------------。

NLP模型的临时测试。此外，回顾一下，给定的*(↪Ll\_1EB↩，k*)-公平性的定义，*k*，表示|*S*|，量化了对输入*x*的所述公平性的信心。

当扰动句子时，利用知识?

虽然所提出的技术能够测试任何NLP任务，但在评估中，我们决定专注于一个广泛使用的NLP任务，即情感分析（SA）。SA已经被很好地商业化，用于日常使用，并不断被报道为受到歧视性输入的影响[Dixon *et al.* , 2018; Kiritchenko and Mohammad, 2018]。

MT-NLP允许用户选择任何合理的敏感属性A作为输入。在评估中，我们将"性别"这个经常被关注的公平因素作为A的敏感属性，我们将其作为未来的一项工作，为MT-NLP配备其他敏感属性，例如"宗教"。

## 扰动

从Large Movie Review测试数据集中的20000个样本句子中，我们的扰动器P找到了17165个"可突变"的句子；可突变句子至少包含一个与人类相关的名词token（见IsPerson）。然后，P总共生成174,183个扰动句作为测试输入，其中有

30 859个是由类比突变产生的，143 324个是由主动突变产生的。因此，对于一个可突变的句子*x*，其扰动集*S*的平均大小为10.2，说明所提出的技术可以揭示出相当数量的任意自然语言句子的测试输入。

图，ConceptNet，是远程托管的，因此，per- turbation时间很大程度上取决于本地网络带宽。因此，扰动时间很大程度上取决于本地网络带宽。尽管如此，我们仍然报告说，我们的扰动需要大约28小时（平均每小时715个样本）。我们在解释整体处理时间时，认为是实用和有前途的。

## 违反测试公平性

表1报告了评估结果。我们使用谷歌、微软和亚马逊提供的三个商业SA服务进行评估。据我们所知，这些商业服务中所部署的NLP模型没有被公开。我们编写了Python脚本来查询这些远程服务并解析它们的输出（JSON格式）。此外，我们还利用大型电影评论训练数据集[Maas*等人*，2011]对两个SA模型进行了lo- cally训练。CNN模型在Keras（2.2.4版）中得到了完善，它有一个卷积层，然后是两个完全连接的层。我们使用IMDB数据的stan- dard预处理模块进行嵌入。至于LR模型，我们使用Scikit-Learn（ver. 0.22.1）中提供的默认设置和BoW嵌入。

虽然有相当数量的扰动句子被生成，但这三种商业NLP服务都有可接受的最大输入长度限制。也就是说，不是每一个sen-

|  |  |  |  |
| --- | --- | --- | --- |
| NLP模式和服务 | 原句 | 烦躁的句子 | 预测变化 |
| Amazon Comprehend API | ....火车去洛杉矶的路上  去找他的叔叔解除诅咒... ...。 | ....火车去洛杉矶的路上  去见他的姑姑以解除诅咒... ...。 | 负数→正数 |
| 微软NLP API | .. . 最喜欢的是O'Reily兄弟. . . | .. . 最喜欢的是O'Reily姐妹. . . | 阳性 → 阴性 |
| 美国有线电视新闻网 | ....一列满载联邦士兵的黄金列车 ...。 | ....一列满载联邦男兵的黄金列车 ...。 | 阳性 → 阴性 |

表2.MT-NLP从商业和本地模型中发现的公平性违规行为。MT-NLP从商业模式和本地模式中发现的公平性违规行为。

100

违规百分比(%)

80

60

40

20

0

24

微软

亚马逊

谷歌

美国有线电视新闻网

LR

LR基地

6 8 10

ε

改变模型的输出。图5报告了对不同NLP模型和服务的评估结果（"LR基线"是由对抗式训练得出的基线，稍后会解释）。百分比的计算方法是，将缓解后的违规数量除以缓解前发现的违规总数（即表1的第三列）。

图5报告了所有NLP模型的可喜结果，尽管它们大概共享不同的模型结构和设置。我们报告说，当*↪Ll\_1EB↩*为0.1时，平均98.0%以上的违规案例可以得到缓解。然而，当*↪Ll\_1EB↩*增加到10时，只有一定数量的违规行为可以减轻。

图5：公平性违规百分比与不同的↪Ll\_1EB↩。

tence可以顺利分析。Ta- ble 1的第二列报告了成功处理的句子数量，第三列报告了触发的公平性违规数量。总的来说，我们报告说，从所有的模型中可以发现相当数量的公平性违规行为。我们将此解释为有希望和直观的；MT-NLP中的MT程序对具体的模型实现是不可知的，而且它在所有这些SA工具中的表现是*一致的*，尽管基础模型存在差异。

与远程商业服务相比，可以从本地mod- els中找到更多的歧视性输入。虽然这些商业工具的实现细节并不封闭，但我们的观察揭示了一个有趣的发现：商业NLP服务似乎会对某些敏感的形容词标记（例如，"girly"）进行*预处理*和修剪，这很可能会导致不必要的和苛刻的歧视。因此，从商业API中发现的大部分歧视性输入都是由类比突变产生的，而类比突变直接扰乱了名词标记。相反，在测试本地模型时，发现了大量由两种突变策略产生的判别性输入。

表2列出了MT-NLP发现的三个典型的判别性输入。第一个例子是在测试亚马逊服务时发现的。MT-NLP识别了与性别相关的名词uncle，并通过类比突变成功推断出相应的al- ternative token，姑姑。同样，第二种情况下，用类比突变将哥哥变为妹妹。最后一个样本呈现了一个在本地CNN模型上主动突变的句子：一个与性别相关的形容词male被插入到原句中，并引发了公平性违规。

## 认证缓解

通过将中立层向NLP模型搭载，我们强制执行*(↪Ll\_1EB↩，k*)公平性保证。在本节中，我们将从不同的设置方面来评估其有效性。在gen- eral中，*↪Ll\_1EB↩*量化了公平性保证的强度，对于足够小的*↪Ll\_1EB↩*，扰动敏感属性不应

微软API(10.6%)和谷歌API(99.3%)中的数据可以是

缓解。在其他情况下，没有任何违规行为得到缓解(因此，在图5中，其相应的百分比为"100%")。

谷歌API的情感分析结果可以根据其返回值（一个浮点数*f*）的符号来判断，如果*f*小于零，则表现为负面情感，反之亦然。另外，如果*f*为零，则该情绪为"中性"。换句话说，对*f进行*小的扰动，例如，0 → 0.001。

足以使其标签由"中性"变为"正性"。

我们检查了谷歌API的公平性违规行为，发现大多数违规行为都是从"正"或"负"误归为"中性"的。因此，即使*↪Ll\_1EB↩*达到10，其他突变的影响可以忽略不计，缓解措施也可以修复大多数错误的预测，通过仍然略微偏离零的*f*。

我们还设计了一个对比组("LR Base"线)，其中LR模型使用原始数据集和我们的pertur- bator P产生的相应突变进行对抗性训练。

23.4%的违规行为，如图5所示)。相比之下，图5中的"LR"数据，当*↪L*l*\_1EB↩*小于1.8时，可以达到更好的缓解率。这一对比表明，我们的缓解技术，在足够小的*↪Ll\_1EB↩*下，与白盒对抗训练相比，可以实现更高的安全保障。

至于成本，由于我们的缓解措施将每个预测*x*扩展为*k*+1个关于{*x*} ∪*S*的预测，每个预测的CPU时间与*k呈*线性关系，如前所述，*k*平均为10.2，这在mod-ern多核CPU上可以很好地并行化。 因此，我们设想经过的实

在现实世界的使用中，时间不应该有很大的变化。总的来说，我们将评估结果解释为有希望的：小*↪Ll\_1EB↩*导致以适度的成本实际缓解公平性违规行为。

# 相关工作

软件测试技术已被应用于测试由机器学习或深度学习模型驱动的AI应用。为此，成熟的测试技术- niques，包括变形测试[Wang*等*，2019；Dwarakanath*等*，2018；Zhang*等*，2018]，differen-。

tial testing [Pei *et al. ,* 2017; Tian *et al. ,*  2018], 和模糊测试[Xie *et al.，*2017；Zhang*等人*，2018]和NLP应用，包括机器翻译[Sun和Zhou，2018；He*等人*，2020；Yan*等人*，2019]和聊天机器人[Bozic和Wotawa，2019]。

=

那么

*y*ˆ*i* =

*y*ˆ*j*

=

(*e↪Ll\_1EB↩* - 1)*yi* + Σ *yn*

(*e↪Ll\_1EB↩* - 1)*yj* + Σ *yn.*

*(e↪Ll\_1EB↩* - 1)*yi* + Σ *yn* - *e↪Ll\_1EB↩*[(*e↪Ll\_1EB↩* - 1)*yj* + Σ *yn*] (*e↪Ll\_1EB↩* - 1)*yj* + Σ *yn*)

+ *e↪Ll\_1EB↩*

AI模型的公平性被特别制定为一个

(1-*e↪Ll\_1EB↩*)(Σ*n)*

*i yn* + *e↪Ll\_1EB↩yi*)

+ *e↪Ll\_1EB↩*≤ *e↪Ll\_1EB↩*

软件财产和软件工程技术分析

niques。到目前为止，突变测试技术被提出来检测NLP模型中的歧视[Galhotra*等人*，2017；Udeshi*等人*，2018]。 Aggarwal*等人*，2019]引入了

(*e↪Ll\_1EB↩* - 1)*yj* + Σ *yn.*

成立。 因此，∀*x，x*′∈{*x* }。∪S ，f(*x*)

≤*f*(*xi*)

(4)

≤*e↪Ll\_1EB↩*。

0 *p f* (*x*′)

*f* (*xj*)

符号执行和局部可解释性来生成内

放在黑箱场景中。尽管如此，现有的研究基本上使用预先定义的映射规则（例如，"男性"→"女性"）来扰动*结构化数据表*（例如，Kag- gle银行营销数据集[Martinez，2018]）。相比之下，这项工作旨在扰动非结构化的自然语言sen- tences，从而广泛地扩展了应用范围。

# 结论

我们已经提出了MT-NLP，一个贯穿始终的自动化工具，以确定和减轻NLP模型的公平性违规行为。我们的评估从商业和本地NLP模型中发现了数以千计的歧视性输入，并在经过认证的保证下一致地平滑模型预测。

# 与人有关的名词词组Ih

我们根据经验将*Ih*定义为：

{"人"、"人"、"人"、"成年人类"、"雇员"}。

# 定理1的证明

在本节中，我们证明定理1中的机制确实是在*(↪Ll\_1EB↩，k*)公平性保证下的。让M表示一个arbi- trary NLP模型，其中X是输入域。我们首先证明M : X → [0*，b*)的情况，然后推广到多维情况。如第2节中介绍的，通过突变

输入x0，我们创建一组语义保存突变*Sp* = {x1*,* - - - *, xk*}。那么根据定理1，我们有*f*

*e↪Ll\_1EB↩* k1

*f* (x0) = *k* + *e↪Ll\_1EB↩* M(x0) + Σ *k* + *e↪Ll\_1EB↩* M(*xi*)(3)

*i*=1

根据定义，我们可以得到一个必然的推论。给定*x，x*′，如果M(*x*)≤M(*x*′)，则*f*(*x*)≤*f*(*x*′)符合。

为了便于表达，给定一组突变

{x0*，*x1*，*--- *，xk*}，让{y0*，*y1*，，，yk*}表示输出。

M和{*y*ˆ0*，y*ˆ1*，，y*ˆ*k*}分别表示*f*的输出，。

另外，在不失一般性的前提下，让*i，j*∈[*k*+1]，其中

[*k*+1] := {0*，*1*, ，k*}。

那么，将不等式反复应用于多维模型，*f*（*x*）≤*e↪Ll\_1EB↩f*（*x*′）成立。

# 参考文献

Aggarwal *et al.* , 2019] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha.Black box fairness testing of machine learning models.In *ACM ESEC/FSE*, pages 625-635.ACM，2019年。

Barocas *et al.* , 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan.*Fairness and Machine Learning*. fairmlbook.org, 2019. [http://www.fairmlbook.org.](http://www.fairmlbook.org/)

Bolukbasi *et al.* , 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai.Man is to computer programmer as woman is to home- maker? debiasing word embeddings.In *NIPS*, pages 4349-4357, 2016.

Bozic and Wotawa, 2019] Josip Bozic and Franz Wotawa.Testing chatbots using metamorphic relations.IFIP- ICTSS, pages 41-55, 2019.

[Chen *et al.* , 1998] Tsong Y Chen, Shing C Cheung, and Shiu Ming Yiu.Metamorphic testing: a new approach for generating next test cases.技术报告，技术报告HKUST-CS98-01，1998.

Cohen *et al.* , 2019] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter.Certified adversarial robustness via random- ized smoothing.In *ICML*, pages 1310-1320, 2019.

Dixon *et al.* , 2018] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman.Measuring and miti- gating unintended bias in text classification.2018.

Dwarakanath *et al.* , 2018] Anurag Dwarakanath, Manish Ahuja, Samarth Sikand, Raghotham M. Rao, R. P. Ja- gadeesh Chandra Bose, Neville Dubash, and Sanjay Pod- der.Identifying implementation bugs in machine learning based image classifiers using metamorphic testing.ISSTA 2018, pages 118-128, 2018.

Ethayarajh *et al.* , 2019] Kawin Ethayarajh, David Duve- naud, and Graeme Hirst.Towards understanding linear word analogies.ACL，2019.

Galhotra *et al.* , 2017] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou.Fairness testing: testing software for discrimination.In *ACM ESEC/FSE*, pages 498-510.ACM, 2017.

Garg *et al.* , 2019] Sahaj Garg, Vincent Perot, Nicole Limti- aco, Ankur Taly, Ed H Chi, and Alex Beutel.Counterfac- tual fairness in text classification through robustness.In *AIES*, pages 219-226.ACM，2019年。

Ge *et al.* , 2018] Tao Ge, Furu Wei, and Ming Zhou.Fluency boost learning and inference for neural grammatical error correction.In *ACL*, pages 1055-1065, 2018.

He *et al.* , 2020] Pinjia He, Clara Meister, and Zhendong Su.Structure-invariant testing for machine translation.ICSE'20，2020.

Huang *et al.* , 2019] Po-Sen Huang, Robert Stanforth, Jo- hannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli.Achiev- ing verified robustness to symbol substitutions via inter- val bound propagation.*arXiv preprint arXiv:1909.01492*, 2019.

Jia *et al.* , 2019] Robin Jia, Aditi Raghunathan, Kerem Go¨ksel, and Percy Liang.Certified robustness to adversar- ial word substitutions.*arXiv preprint arXiv:1909.00986*，2019.

[Kairouz *et al.* , 2016] Peter Kairouz, Sewoong Oh, and Pramod Viswanath.Extremal mechanisms for local dif- ferential privacy.*JMLR*，17（1）：492-542，2016.

Kiritchenko and Mohammad, 2018] Svetlana Kiritchenko and Saif M. Mohammad.Examining gender and race bias in two hundred sentiment analysis systems.*CoRR*，abs/1805.04508，2018.

Kusner *et al.* , 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva.Counterfactual fairness.In

* 1. Guyon，U.V.Luxburg，S.Bengio，H.Wallach，R.Fer- gus，S.Vishwanathan，和R.Garnett，编辑，*NIPS*，4066-4076页。Curran Associates, Inc., 2017.

Lecuyer *et al.* , 2019] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana.Cer- tified robustness to adversarial examples with differential privacy.In *IEEE S&P,* pages 656-672.IEEE，2019。

Maas *et al.* , 2011] Andrew L Maas, Raymond E Daly, Pe- ter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts.Learning word vectors for sentiment analysis.In *ACL*, pages 142-150.Association for Computational Lin- guistics, 2011.

[Manning *et al.* , 2014] Christopher D. Manning, Mihai Sur- deanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky.The Stanford CoreNLP natural language processing toolkit.In *ACL System Demonstrations*, pages 55-60, 2014.

Martinez，2018] Janio Martinez.Kaggle Bank Marketing Dataset，2018.

Mu *et al.* , 2019] J. Mu, P. Liang, and N. Goodman.Shaping visual representation with language for few-shot classifi- cation.*arXiv preprint arXiv:1911.02683*, 2019.

Park *et al.* , 2018] Ji Ho Park, Jamin Shin, and Pascale Fung.Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

Pei *et al.* , 2017] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana.Deepxplore:Automated whitebox testing of deep learning systems.In *SOSP*, pages 1-18.ACM, 2017.

Rhue，2018] Lauren Rhue.Racial influence on automated perceptions of emotions.*Available at SSRN 3281765*, 2018.

Sheng *et al.* , 2019] Emily Sheng, Kai-Wei Chang, Premku- mar Natarajan, and Nanyun Peng.该女子担任保姆。On biases in language generation. *arXiv:1909.01326*, 2019.

Speer *et al.* , 2017] Robert Speer, Joshua Chin, and Cather- ine Havasi.Conceptnet 5.5: An open multilingual graph of general knowledge.In *AAAI*, 2017.

孙和周，2018]孙立群，周志全.Meta- morphic testing for machine translations:MT4MT。ASWEC' 18, pages 96-100, 2018.

Sun *et al.* , 2019] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang.Mitigating gender bias in natural language process- ing:Literature review.*arXiv:1906.08976*，2019.

Tian *et al.* , 2018] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray.DeepTest:Automated testing of deep- neural-network-driven autonomous cars.ICSE '18，2018.

Udeshi *et al.* , 2018] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay.Automated directed fairness test- ing.In *ACM/IEEE ASE*, pages 98-108.ACM, 2018.

Wang *et al.* , 2019] Jingyi Wang, Guoliang Dong, Jun Sun, Xinyu Wang, and Peixin Zhang.Adversarial sample de- tection for deep neural network through model mutation testing.ICSE '19, pages 1245-1256, 2019.

[Warner，1965] Stanley L Warner.随机反应。A survey technique for eliminating evasive answer bias. 美国统计协会杂志，60(309): 63 -69, 1965.*Journal of the American Statistical Association*, 60(309):63-69, 1965.

Xie *et al.* , 2018] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Hongxu Chen, Minhui Xue, Bo Li, Yang Liu, Jian- jun Zhao, Jianxiong Yin, and Simon See.Coverage- guided fuzzing for deep neural networks.*arXiv preprint arXiv:1809.01266*，2018.

严等 ，2019] 严博洋，Brian Yecies，周志全.Metamorphic relations for data validation: a case study of translated text messages.In *2019 IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*, pages 70-75.IEEE，2019.

Zhang *et al.* , 2018] Mengshi Zhang, Yuqun Zhang, Ling- ming Zhang, Cong Liu, and Sarfraz Khurshid.Deep- Road。GAN-based Metamorphic Testing and Input Valida- tion Framework for Autonomous Driving Systems.ASE，2018.

Zhao *et al.* , 2018] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang.核心参考解析中的性别偏见。Evaluation and debiasing methods. *arXiv:1804.06876*, 2018.