基于信息论方法的去偏恶意评论分类探究

TEAM 42

2021 年 12 月 31 日

摘要

恶意评论分类能够自动判别一段评论文本是否带有恶意倾向，有助于社区管理员维护互联网社区环境。由于恶意评论中往往会提及一些敏感身份词，例如 “黑人”、“同性恋” 等等，模型在学习对评论进行分类的时候往往会错误地将身份词与恶意评论建立因果关系，因此需要去偏方法消除模型学习过程中产生的这种偏见。本文使用 BERT 模型，结合基于信息论方法的去偏算法解决恶意评论分类问题。该方法通过建立概率图模型消除神经网络推理中的环境因素。模型在测试集上达到了 93.41% 的准确率。模型实现代码公布在：https://github.com/hiyouga/Toxic_Detection。

团队分工

- **郑耀威**：数据集预处理、TextCNN 模型实现、技术报告的摘要和结论部分撰写。
- **陈克勤**：InvRat 模型实现、技术报告的方法部分撰写。
- **罗哲焱**：BERT 模型实现、技术报告的实验部分撰写。
- **丁雨宽**：超参数调优、技术报告的技术调研部分撰写。
- **张金龙**：数据集分析、技术报告的问题介绍部分撰写。

1 问题介绍

恶意评论分类是自然语言处理问题中经典的篇章级文本分类问题，给定一段文本，该任务旨在判断这段文本是否具有恶意倾向，包括人身攻击、人身侮辱等等。使用机器自动化地分类恶意评论有助于网站管理员维护健康的社区环境。

本问题提供了包含语句文本、标签以及恶意程度评分的训练集，以及只包含语句文本的测试集，训练集包含 1048575 个样本，测试集包含 27074 个样本。语句文本由一句话、多句话或多段构成，且普遍具有口语化、非正规表达的特点，如 “???...You trying to hard...#MAGA” 等。标签中包含了 43 种特征，以及通过投票得出的对应特征值，如 “rating”：approved，“funny”：1 等，其中部分标签内容指出了语句文本中是否包含了可能引起攻击性的词汇，如 “black”，“bisexual” 等，以及具有明显倾向性的评价指标，如 “severe_toxicity”，“insult” 等，但部分特征的残缺值较多。样本语句的恶意程度由 [0,1] 区间的评分值给出，分值越大，表明其恶意程度越强。因此，

有效提取语句文本自身的信息以及标签信息，并在模型中结合使用，将有助于对语句文本的恶意程度进行准确评分。

在本文中，我们采用预训练 BERT 模型 [1] 作为分类器的主体架构，以社区言论文本作为输入，使用给定的标注数据训练分类模型。我们将其看作文本二分类问题，首先使用预训练分词器对文本进行分词，然后输入到 BERT 模型提取文本特征，最后使用 Softmax 层计算该言论为恶意言论的预测概率。

由于恶意言论中往往会包含一些敏感身份词，例如 “黑人”、“同性恋” 或 “犹太人” 等词语。因此模型在训练时可能会学习到偏见，错误地将身份词与恶意言论关联起来，导致模型的预测不可靠。以往研究提出了多种多样的方法以减少模型在学习中产生的偏见，例如域对抗（DAT）[2]、不变解释（InvRat）[3] 等等。

本文采用了不变解释方法 [3] 中的思想，使用基于信息论方法的去偏算法减少模型的偏见。该方法首先对输入文本计算遮罩，以寻找能够合理解释模型推理的片段。然后使用两个不同的分类器来消除非合理解释对模型预测的影响，从而减少模型偏见。

在实验中，我们训练了 BERT 模型和使用了不变解释去偏算法的模型。其中微调后的预训练 BERT 模型在测试集上取得了 93.41% 的准确率。然而使用不变解释去偏算法后的 BERT 模型的效果尚不理想，我们在实验部分对该现象进行了初步的分析和解释。

该工作的主要贡献总结如下：

- 我们分析了恶意评论分类数据集，将其建模为文本二分类问题。
- 我们阐述并实现了基于信息论方法的去偏算法以消除模型在学习中产生的偏见。
- 经过微调的预训练 BERT 模型在测试集上取得了 93.41% 的准确率。

2 技术调研

2.1 恶意评论分类

为了保护社交媒体中的用户免受网络恶意评论的骚扰，近年来，许多基于不同模型的恶意评论系统被提出，用于自动检测互联网社区中的恶意评论。Gaydhani[4] 等人利用支持向量机检测推特上的恶意评论。Bojkovsky 和 Pikuliak[5] 基于对抗学习提出用于恶意评论分类的双向长短期记忆（Bi-LSTM）模型。d’Sa[6] 等人使用 fastText 嵌入和 BERT 嵌入作为 CNN 和 Bi-LSTM 分类器的输入特征，并对预训练的 BERT 模型进行微调。

2.2 深度学习去偏

最初版本的恶意评论分类器对某些语句有明显错误分类的趋势，例如含有某些身份术语的明显无恶意的陈述，“我是一个同性恋”，该类句子被赋予了不合理的高恶意评分，肤色、性别、种族、人口、残疾、宗教信仰等，或者这些因素的复杂组合会极大的影响分类的正确率。

部分研究者试图通过处理训练数据集来消除偏见。例如，Dixon[7] 等人通过添加包含敏感词汇的非恶意评论来平衡数据集。Park[8] 等人结合去偏词嵌入和性别互换数据来减少恶意评论

分类任务中的性别偏见。Badjatiya[9] 等人在训练集中采用了基于多重知识统一化的替换偏差敏感词的策略。

另一些研究者更注重模型的修改和无偏特征的学习。Xia[10] 等人使用对抗训练来减少模型将敏感文本误分类为恶意评论的倾向。Mozafari[11] 等人提出了一种新的加权机制，以减少英语推文中的种族偏见。Vaidya[12] 等人 (2020) 应用了一个带有注意层的多任务学习框架，以防止模型学习特定触发词和恶意标签之间的错误关联。

Chang[3] 等人借助博弈论框架来施加不变性，将 IRM 原理扩展到神经预测，并进一步提出了不变解释（InvRat），一种纳入不变约束的解释模型。Chuang[13] 等人在不变解释模型的基础上，排除输入文本中与恶意标签高度但虚假相关的句法和语义模式，并在推理过程中掩盖这些部分。

本文在先前工作的基础上，通过结合 BERT 和不变解释模型，基于信息论方法的去偏算法解决恶意评论分类问题。

3 方法描述

3.1 基础方法

基础方法将该任务建模为文本分类问题。使用 BERT 模型进行微调来解决。

3.1.1 BERT 模型

BERT 是基于 Transformer 架构的预训练模型 [1]。通过在大型语料库上的进行遮罩语言建模和下一句预测两个任务的无监督预训练，BERT 拥有了强大的泛化能力。从而只需要在下游任务上进行少量的微调迭代即可取得不错的性能。

BERT 由许多层的 Transformer 编码器 [14] 堆叠而成，每个编码器由多头自注意力和前馈网络组成。记 $x^l = \{x_1^l, \ldots, x_N^l\}$ 为第 l 层的特征，则第 $l+1$ 层的特征 x^{l+1} 为：

$$x^{l+1} = \text{FNN}(\text{MultiHead}(x^l)) \tag{1}$$

自注意力可以用来建模序列中元素对该序列中其他元素的依赖关系，结构如图 1 （左）所示。BERT 中使用的自注意力首先通过线性层将原始输入映射到三个不同的空间。即 $[Q, K, V] = [W_q, W_k, W_v] \cdot X$。再使用乘法注意力来计算输出。多头自注意力并行使用多个自注意力，并将其各自的结果进行拼接，作为最终输出。结构如图 1 （右）所示。整体公式如下：

$$\begin{aligned}
\text{MultiHead}(X) &= [\text{head}_1, \ldots, \text{head}_n]W^O \\
\text{head}_i &= \text{Self-Attention}(X) \quad \forall 1 \le i \le n \\
\text{Self-Attention}(X) &= \text{Attention}(W_q X, W_k X, W_v X) \\
\text{Attention}(Q, K, V) &= \text{Softmax}(\frac{QK^T}{\sqrt{d}})V
\end{aligned} \tag{2}$$

前馈网络即由简单的线性层和残差层构成。公式如下，其中 act_fn 为激活函数，通常取 gelu 函数。

$$\text{FNN}(X) = X + W_{f_2}\text{act\_fn}(W_{f_1}X) \tag{3}$$

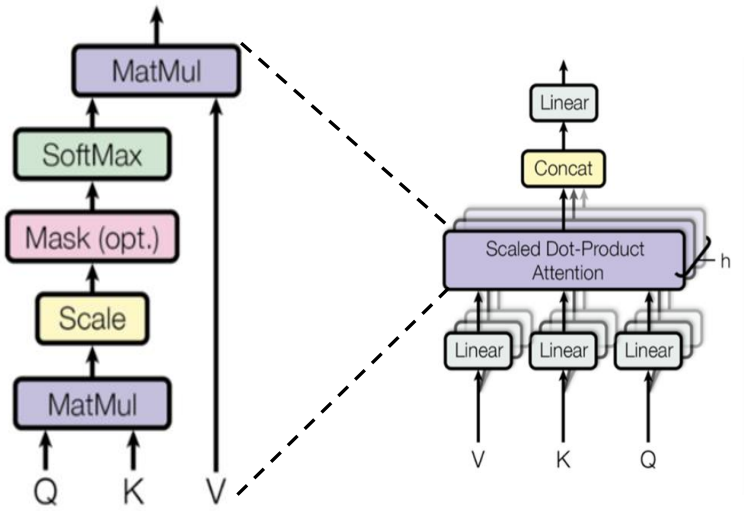

图 1: 多头自注意力结构图 [14]

对于文本分类任务，通常在句子前添加特殊单词 [CLS] 作为句子整体的表示，微调时取 BERT 模型最后一层的 [CLS] 表示进行下游任务的训练。

3.1.2 损失函数

假设训练集中包含着 T 个训练样本 (x_t, y_t)。对于分类问题，通常我们使用交叉熵损失函数优化模型参数，同时也使用权重衰减来正则化模型：

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{T}\sum_{j=1}^{C} y_i^j \log(\hat{y}_i^j) + \lambda \sum_{\theta \in \Theta} \|\theta\|_2^2 \tag{4}$$

其中 y_i^j 代表真实类别，$\hat{y}_i^j$ 代表预测类别，$C = 2$，Θ 对应全部的可训练参数，λ 控制权重衰减正则项对训练过程的影响程度。

3.2 去偏方法

我们尝试基于不变解释 [3] 来构建去偏的恶意言论检测模型。解释是指从输入特征中选取的一个子集，仅通过该特征子集便足够检测出目标。在恶意言论检测任务中，解释可以包含对目标检测有帮助的一切特征，如带有攻击性的句子或词汇，或者指代少数群体的词汇（这部分对检测有帮助往往是因为训练数据有偏）。不变解释是指在不同环境下，该特征子集都可以稳定检测出目标。通过学习不变解释，便可以从有偏数据中过滤出无偏特征，从而得到去偏的恶意言论检测模型。

3.2.1 问题形式化

给定输入输出对 $(\boldsymbol{X}, Y)$，一段语句 $\boldsymbol{X}$，可以被分为 X_1, X_2, X_3 三部分。如图 2 所示。其中 X_1 为影响检测结果 Y 的特征，X_2 为被 Y 直接影响的特征，X_3 为其他特征。X_1, X_2, X_3 都有可能和 Y 高度相关，但只有 X_1 才是合理的解释。为分离出 X_1，不变解释引入了环境变量 E，

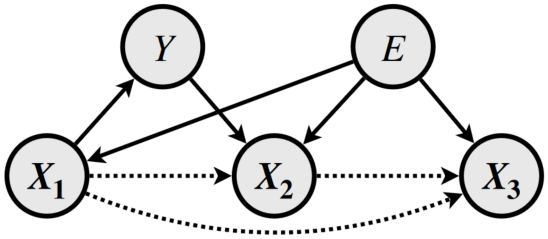

图 2: 不变解释示意图 [3]

环境变量 E 影响着 X 的先验分布，从而导致从 X_2, X_3 预测 Y 的能力在不同环境下发生变化，同时从 X_1 预测的能力保持不变。即 Y 和 E 关于 X_2, x_3 条件不独立, 关于 X_1 条件独立。为过滤出合理的解释 X_1，目标函数可形式化为：

$$\begin{aligned} \max_{m \in S} \quad & I(Y; \boldsymbol{Z}) \\ \text{s.t.} \quad & \boldsymbol{Z} = m \odot \boldsymbol{X} \\ & Y \perp E \mid \boldsymbol{Z} \end{aligned} \tag{5}$$

其中 $I(Y; \boldsymbol{Z})$ 表示 Y 和 $\boldsymbol{Z}$ 的互信息。S 为遮罩空间，包含预设的所有可能的遮罩方案，m 是指输入特征 $\boldsymbol{X}$ 上的一个遮罩，$\odot$ 表示按位乘运算。根据 D 分离定理，当且仅当 $\boldsymbol{Z} = [X_1, 0, 0]$ 时，有 $Y \perp E \mid \boldsymbol{Z}$。

3.2.2 模型框架

为解决式最优化问题 (5)，不变解释（InvRat）方法 [3] 引入训练框架，如图 3 所示。该框架由三部分构成，分别是 1) 生成解释 $\boldsymbol{Z}$ 的生成器 $g(\boldsymbol{X})$；2) 环境无关的检测器 $f_i(\boldsymbol{Z})$；3) 环境相关的检测器 $f_e(\boldsymbol{Z}, E)$。其中检测器 $f_i(\boldsymbol{Z})$，$f_e(\boldsymbol{Z}, E)$ 都用来预测 Y，唯一的不同是是否可以访问环境 E。记 $\mathcal{L}(Y; f)$ 为单个样例上的交叉熵损失函数。则检测器的训练目标分布为：

$$\begin{aligned} \mathcal{L}_i^* &= \min_{f_i(\cdot)} \mathbb{E}[\mathcal{L}(Y; f_i(\boldsymbol{Z}))] \\ \mathcal{L}_e^* &= \min_{f_e(\cdot,\cdot)} \mathbb{E}[\mathcal{L}(Y; f_e(\boldsymbol{Z}; E))] \end{aligned} \tag{6}$$

生成器通过对输入特征 X 进行遮罩生成解释 Z。生成器同样参与最小化环境无关预测器的训练 $\mathcal{L}_i^*$。另外，还有一项用于近似约束 $Y \perp E \mid \boldsymbol{Z}$ 的正则项 $\lambda_{\text{diff}}\, h(\mathcal{L}_i^* - \mathcal{L}_e^*)$，其中 $\lambda_{\text{diff}} > 0$ 为正则项系数，$h(t)$ 为单调增的凸函数。关于该正则项的理论分析，请参阅原论文 [3]。

3.2.3 解释提取

最优化问题 (5) 中的遮罩空间 S，需要满足稀疏性约束和连续性约束。稀疏性约束要求生成的解释 $\boldsymbol{Z}$ 只包含少部分特征；连续性约束要求所截取的特征片段要尽可能连续。这符合人类理解自然语言的特点：一段话是否有恶意，仅仅通过少部分单词，语句就能反应出；相比于任意选取的单词，词组和语句片段能保留更多结构信息，更容易理解。据此，使用两个正则项 [15]：

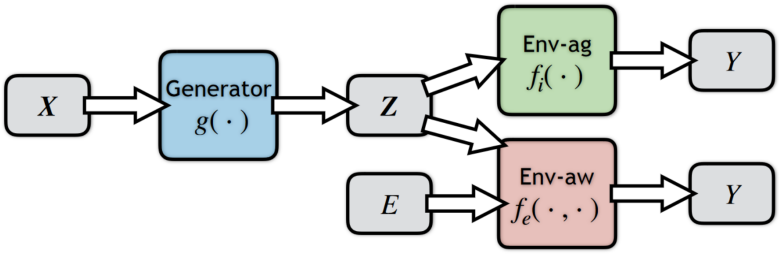

图 3: 不变解释框架图 [3]

$$\lambda_{\text{sparsity}}\,\mathbb{E}\left[\left|\frac{1}{N}\|m\|_1-\alpha\right|\right]+\lambda_{\text{continuity}}\,\mathbb{E}\left[\sum_{n=2}^{N}|m_n-m_{n-1}|\right] \tag{7}$$

其中，$\lambda_{\text{sparsity}},\lambda_{\text{continuity}}>0$ 为正则项系数，α 为所生成遮罩 m 的目标稀疏比例。

3.2.4 损失函数

以上多个目标函数整体上还可以表述成最小化最大值问题的形式：

$$\begin{aligned}\min_{g(\cdot),f_i(\cdot)}\max_{f_e(\cdot)}\quad & \mathcal{L}_i(g,f_i)+\lambda_{\text{diff}}\,h(\mathcal{L}_i(g,f_i)-\mathcal{L}_e(g,f_e))\\ &+\lambda_{\text{sparsity}}\,\mathbb{E}\left[\left|\frac{1}{N}\|m\|_1-\alpha\right|\right]\\ &+\lambda_{\text{continuity}}\,\mathbb{E}\left[\sum_{n=2}^{N}|m_n-m_{n-1}|\right]\end{aligned} \tag{8}$$

其中

$$\begin{aligned}\mathcal{L}_i\left(g,f_i\right)&=\mathbb{E}\left[\mathcal{L}\left(Y;f_i(\boldsymbol{Z})\right)\right]\\ \mathcal{L}_e\left(g,f_e\right)&=\mathbb{E}\left[\mathcal{L}\left(Y;f_e(\boldsymbol{Z},E)\right)\right]\\ \boldsymbol{Z}&=m\odot\boldsymbol{X}\\ m&=g(\boldsymbol{X})\end{aligned} \tag{9}$$

4 模型评估和分析

4.1 数据集分析

我们在给定的训练集和测试集上进行实验，训练集中的每条数据包含评论 ID、评论内容与攻击性评分，测试集则只包含评论 ID 和评论内容。每句评论内容是由英文词组组成的句子。我们将攻击性评分大于等于 0.5 的评论视为攻击性评论，模型需要预测的是一句话具有攻击性的概率。表 1 展示了数据集的统计信息。可以看到攻击性的评论只占全部样本数的约 8%，是一个非常不平衡的数据集。

进一步地，我们对含有特定词汇的语句进行了统计，图 4 展示了包含不同词汇的攻击性评论个数。不同的评论背景下，攻击性评论占比不同，例如包含 “black” 的评论中有接近三分之一

	样本数	最大单词个数	攻击性评论数
训练集	1777800	317	142178
测试集	27074	202	-

表 1: 数据集统计信息

都含有攻击性，并且在这些背景下，有攻击性的比率都高于训练集中所有评论中恶意评论的比率 8%。

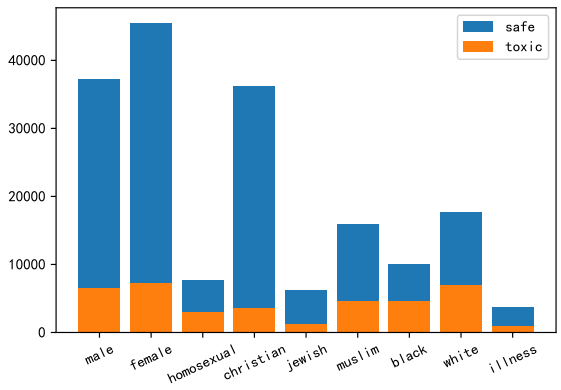

图 4: 不同背景下的攻击性评论数

4.2 评价指标

对于模型输出的概率，使用 Bias AUC 评判模型的好坏，Bias AUC 是一个改进版的 AUC 指标，用于纠正可能存在的模型偏差，达到考察模型去偏能力的作用。

Bias AUC 包含 4 项不同的 AUC，每个 AUC 使用测试集不同的部分进行评价：

- **全体 AUC**：在所有数据上计算 AUC。
- **子集 AUC**：选取出包含特定词汇（如 “male”, “female” 等）的评论计算 AUC。
- **BPSN AUC**：对于特定的词汇，在非恶意评论中选出带有这些词的评论，在恶意评论中选取**不**包含这些词汇的评论计算 AUC。
- **BNSP AUC**：对于特定的词汇，在非恶意评论中选出**不**带有这些词的评论，在恶意评论中选取包含这些词汇的评论计算 AUC。

最终指标 Bias AUC 由这几项加权平均得到：

$$\begin{aligned} Bias\_AUC = &\frac{1}{4} Overall\_AUC + \frac{1}{4}\left(\frac{1}{|G|}\sum_{g\in G} Subgroup\_AUC_g^{-5}\right)^{-\frac{1}{5}} \\ &+ \frac{1}{4}\left(\frac{1}{|G|}\sum_{g\in G} BPSN\_AUC_g^{-5}\right)^{-\frac{1}{5}} + \frac{1}{4}\left(\frac{1}{|G|}\sum_{g\in G} BNSP\_AUC_g^{-5}\right)^{-\frac{1}{5}} \end{aligned} \tag{10}$$

4.3 实现细节

4.3.1 基础方法

将文本分词以后输入 BERT 模型，使用 [CLS] 标签的输出，通过线性分类器进行分类。设定学习率为 10^{-5}，权重衰减为 10^{-5}。使用 Adam 优化器进行优化。分类器特征随机丢弃（Dropout）的概率设为 0.1。我们在全数据集上每 300 次更新进行一次测试，取验证集上表现最好的模型作为最终的模型，得到测试集上的预测结果。

4.3.2 去偏方法

综合考虑模型性能以及算力需求，在不变解释的框架中，分别使用三个参数不共享的 DistilRoBERTa 模型 [16] 作为生成器 $g(\boldsymbol{X})$ 和检测器 $f_i(\boldsymbol{Z}), f_e(\boldsymbol{Z}, E)$ 的骨干模型。对于环境相关的检测器 $f_e(\boldsymbol{Z}, E)$，环境嵌入依次和该样本中所有单词嵌入相加，再输入到之后的骨干模型中。对于生成器 $g(\boldsymbol{X})$ 得到的遮罩 m，强制第一位为 1，保证表征句子特征的特殊单词 [CLS] 在下游的检测器中。

由于 train_extra.json 中类别的缺失较多，关于环境变量的生成，参考 Chuang 等人 [13] 基于规则的工作，采用将基于规则和 train_extra.json 标注相结合的方法构建。使用正则表达式，匹配句子中特定的单词出现，将句子分为“有不针对少数群体的恶意单词”，“有针对少数群体的恶意单词”，“有针对少数群体的单词（但无恶意）”和“其他”共四类。每一类的嵌入保证和单词嵌入维数相同，并使用随机初始化。

所有句子统一填充或截取到 256 个单词。

4.4 超参数调优

由于训练集过于庞大，而预训练模型训练过于耗时，每次利用全数据进行训练并进行参数调整会使得开销增大，因此我们先在基线模型上划分出不同量的训练数据进行测试，并利用少量数据对我们的模型调优。具体来说，TextCNN 模型上不同数据量的表现如表 2 所示，其中验证集通过从训练数据中划出 5% 或 10% 留出得到，在数据量较大时选择留出比例更小的验证集。最终，我们用 10^5 条数据进行训练并调优我们的模型，从中我们划出 5% 作为验证集。在最终训练时使用全部的训练数据，使用 2% 的数据进行验证。

训练数据条数	验证集 Bias AUC	测试集 Bias AUC
10^5	89.22	88.83
5×10^5	90.10	89.94
1.78×10^6	91.29	91.32

表 2: 不同量训练数据下 TextCNN 的表现

4.5 结果对比

经过模型训练和超参数调整，我们对比了不同方法的效果，结果如表 3 所示，评测指标采用 Bias AUC。最终我们发现经过微调后的预训练 BERT 模型取得了最优的效果，而加入不变解释框架的去偏算法的性能不太理想，在测试集的准确率降低了 2 个百分点。可能的原因是敏感身份词的标注缺失值较多，模型无法很好地消除环境因素。最终我们的模型在比赛工作站上取得了第 6 名的成绩，如图 5 所示。

模型	Bias AUC
TextCNN	91.32
AdvBERT	92.96
InvRat	91.20
DistilBERT	92.08
BERT	**93.41**

表 3: 模型表现对比

图 5: 比赛工作站排名截图

5 结论

在这篇文章中，我们研究了使用深度神经网络模型解决恶意评论分类的问题。首先我们对给定的训练数据集进行了详细的分析，观察了每个身份词与恶意评论的关系，也就是模型可能学习到的偏见信息。接着我们将恶意评论分类问题建模为文本二分类任务，使用预训练的 BERT 模型提取文本特征，计算文本为恶意评论的概率。为了消除模型在训练中产生的偏见，我们引入了基于信息论的去偏算法，该方法通过训练两个不同的分类器来寻找输入样本中的无偏特征。在实验中，我们使用给定的数据集对预训练 BERT 模型进行微调，使用少量数据筛选最佳的超参数，最终模型在测试集上取得了 93.41% 的准确率，在比赛工作站上取得了第 6 名的成绩。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[3] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020.

[4] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[5] Michal Bojkovskỳ and Matúš Pikuliak. Stufiit at semeval-2019 task 5: Multilingual hate speech detection on twitter with muse and elmo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 464–468, 2019.

[6] Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:* "*Organization of Knowledge and Advanced Technologies*" *(OCTA)*, pages 1–5. IEEE, 2020.

[7] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

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

[9] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59, 2019.

[10] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020.

[11] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[12] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

[13] Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. Mitigating biases in toxic language detection through invariant rationalization. *arXiv preprint arXiv:2106.07240*, 2021.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[15] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. A game theoretic approach to class-wise selective rationalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10055–10065, 2019.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.