# 预测模型

本部分旨在建立一个网络暴力语言预测模型，以识别网络暴力语言并进行细粒度分类。本部分数据集采用TOXICN数据集，该数据集是由Junyu Lu, Bo Xu(2023)[[1]](#footnote-1)提出的中文互联网的网络暴力语言细粒度分类数据集。预测模型在预训练大语言模型(PLM)如Bert基础上，通过TKE、侮辱词典等辅助手段加强语言模型对于网络暴力语言的识别能力。

## 1.1 数据集

### 1.1.1 数据集建立

我们使用了TOXICN数据集，该数据集抓取的数据范围限定在几个敏感话题下，包括性别、种族、地域黑和LGBTQ，这些话题在互联网上很容易引起争论。其中列出了每个话题的一些关键词，并利用这些关键词提取出了总计15442条 没有回复的评论。删除了文本过于简短而没有实际语义的样本，比如只由词形变化和助词组成的短语。删除了包括重复的样本和不相关的广告等数据。最后，保留了12011条评论。

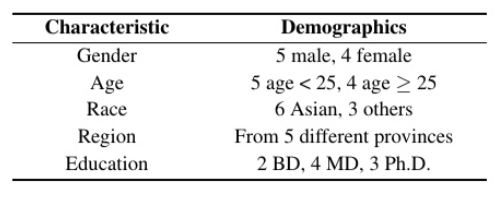
以下是数据集的部分样本示例。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Platform** | **Topic** | **Content** | **Toxic** | **Toxic One Hot** | **Toxic**  **Type** | **Toxic Type One Hot** | **Expression** | **Expression**  **One Hot** | **Target** | **Length** | **Num**  **Attacked** |
| zhihu | region | 就凭你的北京户口我就感觉你比我高级五个阶级，我是河北人😆级五个阶级，我是河北人 | 0 | [1, 0] | 0 | [1, 0, 0] | 0 | [1, 0, 0, 0] | [0, 0, 0, 0, 0, 1] | 31 | 0 |
| tieba | race | 抓虫大  战，人人  有责 | 1 | [0, 1] | 2 | [0, 0, 1] | 2 | [0, 0, 1, 0] | [0, 0, 0, 1, 0, 0] | 9 | 1 |
| tieba | race | 好久都没有见过那么干净的白人班级了，一个黑乐色都没有 | 1 | [0, 1] | 2 | [0, 0, 1] | 1 | [0, 1, 0, 0] | [0, 0, 0, 1, 0, 0] | 26 | 1 |

其中platform指数据来源平台，topic指该评论出自什么主题的帖子，content即为文本内容，toxic即为评判是否网络暴力的二分类标注，toxic\_type表示其为对言论的分类一般攻击性语言和仇恨性言论。expression表示表达方式，target表示攻击的目标群体，length表示文本内容长度，num\_attacked表示攻击目标个数。 对于一般攻击性语言和仇恨性言论的区别，根据Waseem and Hovy[[2]](#footnote-2)(2016)和Fortuna and Nunes[[3]](#footnote-3)(2018)，以下是几个识别仇恨言论的标准:1)攻击特定群体，或2)煽动他人仇恨少数群体，或3)基 刻板印象和扭曲事实对少数群体制造偏见、排斥或厌恶，或4)使用讽刺或幽默来嘲笑群体，尽管发布者可能并非恶意。相比之下， 一般的攻击性语言并没有对具有特定社会属性的目标进行侮辱(David-son et al.2017)[[4]](#footnote-4)。

### 1.1.2 偏执处理

标注者的主观偏见会对数据集的质量产生负 面影响(Waseem和Hovy, 2016)。因此，在标 注的设计和构建过程中，mit-igate这些偏差 是很重要的。为此，我们采取以下措施:我们 首先保证标注者在背景信息方面的多样性， 包括性别、年龄、种族、地区和研究。所有参与者均为语言学专业，并经过系统训练。 标注者的人口统计数据如表所示。然后， 我们对爬取的帖子中包含的网络暴力内容进行渐 进式分析，初步确定各种粒度的标注规则。 经过几次迭代的小规模标注测试和对边缘情况的讨论，最终确定了标准。



## 1.2 模型建立

主要使用了Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks中的模型与方法。

### 1.2.1 主要方法和思路

**分层分类法（MONITOR TOXIC FRAME）：**

**目的:** 区分仇恨言论和一般冒犯性语言，并分析网络暴力语言表达方式。

**层级：**

网络暴力识别： 二元分类，用于确定评论是否包含网络暴力语言。

网络暴力类型区分： 区分一般冒犯性语言和仇恨言论。

目标群体和表达类型检测： 识别目标群体（性别歧视、种族歧视、地域偏见和反LGBTQ）以及表达类型（显性、隐性或报告）。

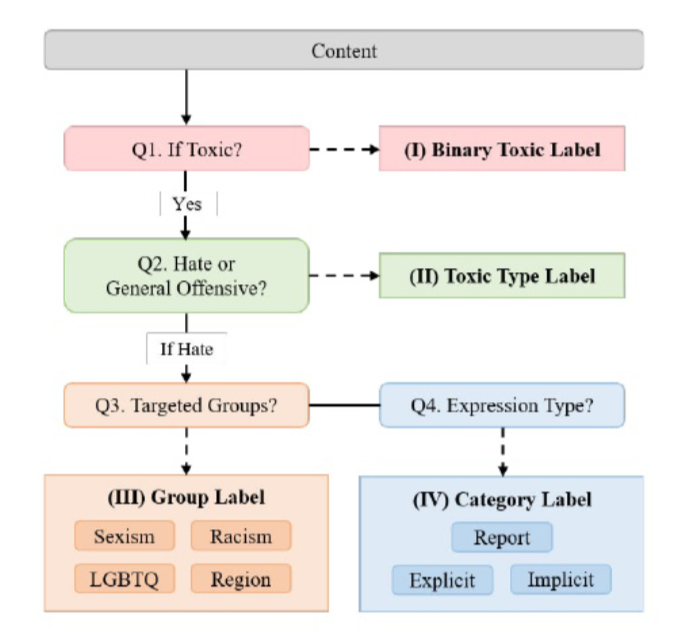
**侮辱词典：**

**目的：** 包含隐性脏话，这对于检测微妙的网络暴力语言至关重要。

**内容：** 包含显性脏话和隐性网络暴力术语，如讽刺性隐喻。

**网络暴力知识增强（TKE）：**

**描述：** 一个基准，结合词汇特征来增强文本表示，改善网络暴力语言的检测。

**评估：** 使用多个基准结合TKE，证明其在检测网络暴力语言方面的有效性。

### 1.2.2深度学习基准模型

我们使用多个基准模型作为编码器进行实验，包括BiLSTM和预训练语言模型（PLM），如BERT和RoBERTa。 这里我们介绍实验的基准模型。以下几个现有的基准模型被用作编码器。我们使用一个全连接层作为几个子任务的分类器。

BiLSTM。该方法采用腾讯AI Lab Embedding6的静态词向量，具有200维特征， 并使BiLSTM整合上下文信息。

BERT。两个最常用的基于中文的transform文本模型, bert- basedchinese7 和 robertabasedchinese8，被用作基准模型。在实验中，编码器的池化输出被用作连接分类器的输入。

### 1.2.3侮辱词典

根据攻击对象的不同，将标注过程中构建的侮辱词典分为五类。该词典包括性别歧视、 种族主义、地域偏见、反lbgtq以及一般脏话， 一般脏话指的是那些可以用来冒犯任何群体的脏话。 最终的词典收录了1032个词条，不仅包括露骨的粗俗词汇，也包括含蓄的侮辱性词汇。

此外，由于互联网每年都会产生大量新的辱骂词，为了规避过滤机制，网民改变原有的辱骂词，创造出字形和发音相似的新变体(陈2012;Zhang, 2010)，分别被称为变形和谐音。在一些脏话中，汉字有时会被其他语言取代形成混合词或缩写。还有一些辱骂词是引入有言外之意的术语，包括隐喻和反讽的使用(陈，2012)[[5]](#footnote-5)。除此之外，一些含有特定的偏见的外来词也被用于隐性的网络暴力评论中(Shi, 2010)[[6]](#footnote-6)。与基于表面特征的变体相比， 这些具有深层语义的术语必须借助背景知识进行检测。

我们进一步从提出的侮辱词中分析网络脏话的派生规则。我们从表面特征和实际含义两个方面对它们进行了简要总结。以下是一些派生方式的总结：

变形。由于汉字是象形文字，通过与单个 汉字的分离和结合，赋予汉字包含特定情感的 意义(陈，2012)。一个例子是“默”(意为“沉 默”)，其字形由“黑”(意为“黑”)和“犬” (意为“狗”)组成，含蓄地表达了对黑人群体 的厌恶。

褒词贬用。褒义词有时是讽刺地使用来达到侮辱的效果，这往往体现在带有新含义的旧词 上(Fortuna and Nunes, 2018)。就像“仙女” (意为“仙女”)，原本温柔善良的形象被暗含 为粗鲁无礼的“泼妇”。

缩写。例如“txl”，每个字母分别是 “同”、“性”和“恋”的发音首字母，意思是“gay”。

语种混合。为了强调语气，在中文网络平台的文本中广泛混合了非中文语言，比如英文和表情符号(Li etal.， 2020)[[7]](#footnote-7)。比如脏话“ni哥”(意为“ni哥”)， 和“nagger”的读音相同。

外来词。一些外来词弥漫着某种网络暴力的文化内涵(Shi, 2010)。因此，需要背景信息来了解这些术语的实际语义。一个例子是 “凯勒奇”，指的是反犹的卡勒吉计划，被用作一个煽动性的术语。

### TKE方法描述

TKE方法旨在通过结合词汇特征来增强文本表示，改善网络暴力语言的检测能力。具体步骤如下：

**词汇嵌入：** 对于给定的句子 ，每个词元 被嵌入为一个维向量 。

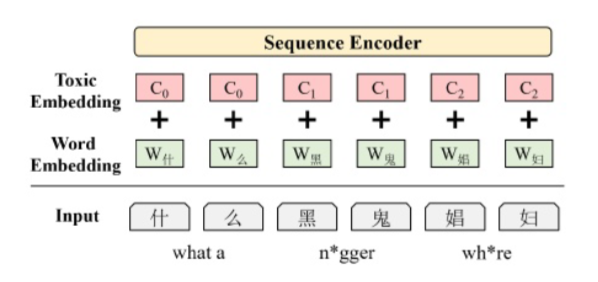
**网络暴力嵌入设计：** 利用n-gram方法确定 是否为侮辱词的子词，并根据攻击的群体类别进行标注。随机初始化群体类别表示 ，其中表示非网络暴力词， 表示侮辱词典的类别数。在此工作中，。

**网络暴力嵌入定义：** 对于 ，其网络暴力嵌入 的定义如下：

，如果 是非网络暴力词。

，如果 属于第 类。

**增强表示：** 将网络暴力嵌入和词汇嵌入 w\_i 进行逐元素加法得到增强表示，其中 是一个控制网络暴力知识摄取的权重系数。最终的句子嵌入表示为 。



## 1.3 预测结果

### 1.3.1实验结果解释

我们采用了广泛使用的加权精度(P)、召回率(R) 和F1-score (F1)指标来评估模型的性能。利用 加权交叉熵来解决类别不平衡问题，优化器是 AdamW。在训练阶段应用了早期停止机制。所有样本被分成训练集和测试集，比例 为8:2。我们对超参进行微调，并在测试集上保 留表现最好的模型和超参数，通过改变随机种子来减少误差，重复5次相同的实验。

### 1.3.2训练集及验证集结果

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Epoch** | **Train**  **F1** | **Train**  **All F1** | **Train**  **Precision** | **Train**  **Recall** | **Train**  **Loss** | **Eval**  **F1** | **Eval All F1** | **Eval**  **Precision** | **Eval**  **Recall** | **Eval**  **Loss** |
| 0 | 0.7488 | [0.7024,0.7788]  0.7788] | 0.7579 | 0.7463 | 0.5268 | 0.7985 | [0.7787,  0.8142] | 0.8005 | 0.7980 | 0.4629 |
| 1 | 0.8538 | [0.8393,  0.8657] | 0.8540 | 0.8536 | 0.3627 | 0.8036 | [0.7888,  0.8161] | 0.8040 | 0.8034 | 0.4498 |
| 2 | 0.9045 | [0.8953,  0.9122] | 0.9047 | 0.9045 | 0.2624 | 0.8140 | [0.7870,  0.8326] | 0.8218 | 0.8125 | 0.4567 |
| 3 | 0.9414 | [0.9358,  0.9460] | 0.9415 | 0.9414 | 0.1786 | 0.8052 | [0.7904,  0.8178] | 0.8057 | 0.8051 | 0.5055 |
| 4 | 0.9672 | [0.9642,  0.9697] | 0.9672 | 0.9672 | 0.1122 | 0.8070 | [0.7621,  0.8319] | 0.8321 | 0.8030 | 0.6130 |

### 1.3.3测试集结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Test F1** | **Test All F1** | **Test Precision** | **Test Recall** | **Test Loss** |
| Test | 0.8129 | [0.7880, 0.8306] | 0.8189 | 0.8117 | 0.4619 |

## 1.4 结果与讨论

1. **实验结果：** 结果表明，深度学习方法相比于在线API如百度文本审核，具有更好的性能。这可能是因为在线API的过滤机制较为简单。

2. **不同子任务的性能：** 实验结果显示，TKE在不同粒度的网络暴力语言检测任务中提高了模型的性能，尤其在隐性网络暴力表达的检测中效果显著。

3. **错误分析：** 通过对所有模型误分类样本的人工检查，识别出两种主要的错误类型：

**类型I错误（假阴性）：** 模型将网络暴力句子分类为非网络暴力，通常是由于缺乏语义层面的背景信息。

**类型II错误（假阳性）：** 模型将非网络暴力句子分类为网络暴力，通常是由于标记中的脏话或少数群体代词引起的误判。

## 1.5 参考文献

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, Hongfei Lin School of Computer Science and Technology, Dalian University of Technology, China.2023.Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech

detection and the problem of offensive language.

1. Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, Hongfei Lin School of Computer Science and Technology, Dalian University of Technology, China.2023.Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks [↑](#footnote-ref-1)
2. Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. [↑](#footnote-ref-2)
3. Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. [↑](#footnote-ref-3)
4. Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech

   detection and the problem of offensive language. [↑](#footnote-ref-4)
5. Wangdao Chen. 2012. Rhetoric introduction. Fudan University Press. [↑](#footnote-ref-5)
6. Chunhong Shi. 2010. Web language as a language variety and a linguistic issue. Applied Linguistics,

   (3):70–80. [↑](#footnote-ref-6)
7. Bin Li, Yan Dou, Yingting Cui, and Yuqi Sheng. 2020. Swearwords reinterpreted: New variants and uses by young chinese netizens on social media platforms.

   Pragmatics, 30(3):381–404. [↑](#footnote-ref-7)