

# 【外部】飞书安全-内容安全检测算法

AI 速览

本文讨论了飞书安全的内容安全检测算法项目，该项目结合文本分析与图像识别技术，进行情感识别、动物检测模型训练及大模型图文识别应用，介绍了需求、交付...

## 一、项目背景

本项目旨在结合文本分析与图像识别技术，针对特定图文内容进行识别检测，实践情感识别、动物检测两个模型的训练过程，实验大模型在图文识别任务中的应用。















## 二、需求说明

快速扫描：训练文本、图像模型，实现特定信息的判别任务。

精细扫描：使用大模型实现更精细的判别。

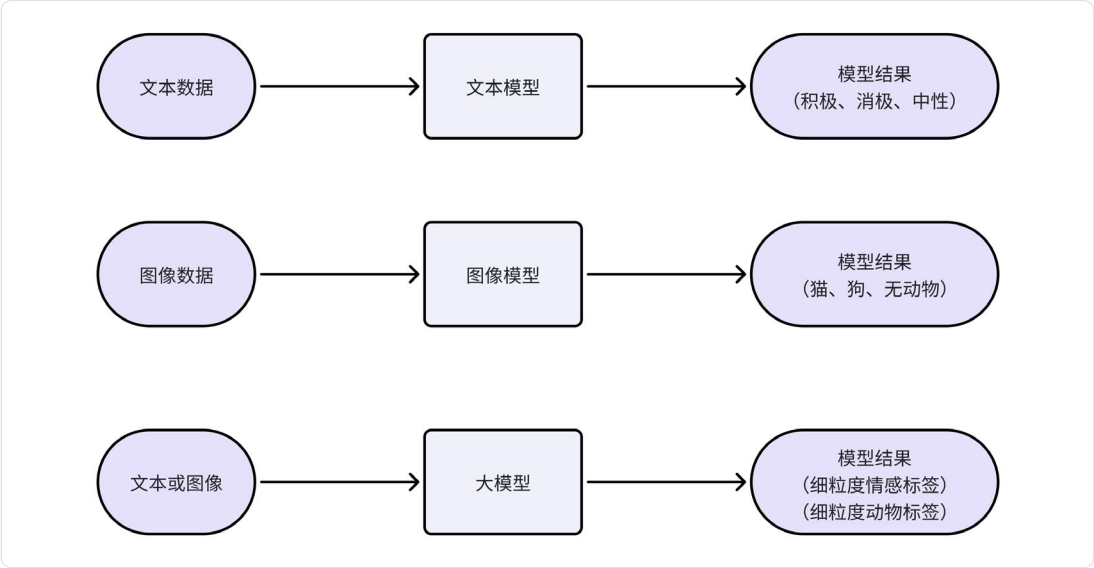
目标	需求详情
任务1：文本分类-情感识别模型快速扫描	<div>1. 训练模型，实现文本情感识别（积极、消极、中性）</div> <div>2. 保证CPU（4C8G）模型推理速度 &lt; 500ms（文本长度512字符）</div> <div>3. 本地测试集F1达标，并在三次提交非公开测试集上取得更好效果。</div>
任务2：图像分类-动物检测模型快速扫描	<div>1. 训练模型，实现动物检测（猫、狗、无动物）</div> <div>2. 保证CPU（4C8G）模型推理速度 &lt; 500ms（256x256像素）</div> <div>3. 本地测试集F1达标，并在三次提交非公开测试集上取得更好效果。</div>
任务3：大模型图文精细扫描	<div>1. 细粒度情绪识别，使用大模型判别细粒度情绪类型（高兴、悲伤、愤怒、平静、阴阳怪气）</div> <div>2. 细粒度图像检测，使用大模型判别图像内动物标签和位置框。</div>

### 任务数据集

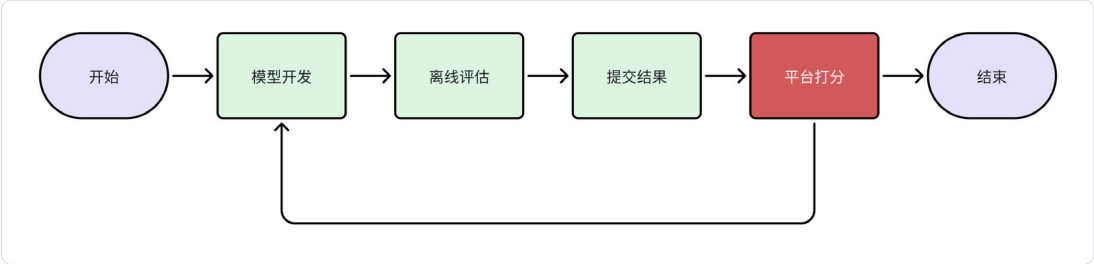
	文本任务-情感识别	图像任务-猫狗识别
基础训练集（公开）	<div>标签：积极、消极、中性</div> <div><div>base_train.csv.zip</div><div>3.55MB</div><div></div></div>	<div>标签：猫、狗、其他</div> <div><div>image_train.zip</div><div>642.38MB</div><div></div></div>
基础验证集（公开）	<div>标签：积极、消极、中性</div> <div><div>base_val.csv.zip</div><div>1.78MB</div><div></div></div>	<div>标签：猫、狗、其他</div> <div><div>image_val.zip</div><div>63.95MB</div><div></div></div>
大模型验证集（公开）	<div>标签：积极、愤怒、悲伤、恐惧、惊奇、中性</div> <div><div>adv_val.csv.zip</div><div>1.07MB</div><div></div></div>	<div>标签：猫/狗；位置框</div> <div><div>val.zip</div><div>22.29MB</div><div></div></div> <div><div>image_label_val.csv</div><div>60.39KB</div><div></div></div>

### 基本功能说明

#### 模型功能



快速扫描模型评测（文本、图像）



可选功能

- 1. 模型结果可视化：具体见评分标准

三、项目交付

交付清单

- 文档类
  - 给出详细的算法方案设计
- 代码
  - 项目源代码，需提供模型预测执行脚本验证测试集指标：
    - text\_predict.py（该脚本需读取目标文本文件的所有文本预测文本违规概率，生成text\_predict.csv文件，该文件包含两列，一列text，一列score）

```
文本任务预测脚本（text_predict.py 示范代码）

1  import pandas as pd
2  import torch
3  import csv
4
5  # load your model
6  model = Model().cuda()
7  id2label = {0: '积极', 1: '消极', 2: '中性'}
8
9  df = pd.read_csv('base_val.csv', dtype=str, keep_default_na=False)
10 res = {'text': [], 'label': []}
11 for _, item in df.iterrows():
12     text = df['text']
13     scores = model(text)
14     score, label = torch.max(scores, dim=-1)
15     res['text'].append(df['text'])
16     res['label'].append(id2label[label])
17
18 text_df = pd.DataFrame(data=res, dtype=str)
```

```
19 text_df.to_csv('text_predict.csv', index=False, quoting=csv.QUOTE_ALL)
```

- image\_predict.py (该脚本需读取目标文件夹下的所有图片预测图片违规概率，生成image\_predict.csv文件，该文件包含两列，一列image\_path，一列score)

```
图像任务预测脚本 (image_predict.py 示范代码)

1  import pandas as pd
2  import torch
3  import csv
4  import glob
5  from PIL import Image
6
7  # load your model
8  model = Model().cuda()
9  id2label = {0: 'cat', 1: 'dog', 2: 'other'}
10
11 res = {'image_path': [], 'label': []}
12 for image_path in glob.glob('data/test/*'):
13     scores = model(Image.open(image_path))
14     score, label = torch.max(scores, dim=-1)
15     res['image_path'].append(image_path)
16     res['label'].append(id2label[label])
17
18 image_df = pd.DataFrame(data=res, dtype=str)
19 image_df.to_csv('res.csv', index=False, quoting=csv.QUOTE_ALL)
```

- 汇报演示 (非必须，如有更好)
  - 需演示正常图片、文本等内容的表现以及违规内容的表现

考核标准

评价标准 (合格-60分)	评价标准 (优秀-80分)	评价标准 (超出期望-90分以上)
<ul style="list-style-type: none"><li>文本模型<ul style="list-style-type: none"><li>F1 &gt; 0.8</li><li>推理延时 &lt; 500ms</li></ul></li><li>图像模型<ul style="list-style-type: none"><li>F1 &gt; 0.8</li><li>推理延时 &lt; 500ms</li></ul></li><li>大模型<ul style="list-style-type: none"><li>实现文本、图片的分类任务，产生可读的分析结果</li></ul></li></ul>	<ul style="list-style-type: none"><li>文本模型<ul style="list-style-type: none"><li>F1 &gt; 0.9</li></ul></li><li>图像模型<ul style="list-style-type: none"><li>F1 &gt; 0.9</li></ul></li><li>大模型<ul style="list-style-type: none"><li>结果格式标准且稳定，可用于自动化任务</li><li>能够正确判别细粒度情感</li><li>能够正确识别动物类别和位置框</li></ul></li></ul>	<ul style="list-style-type: none"><li>文本模型<ul style="list-style-type: none"><li>模型结果可视化，输出文本中与结果相关的片段</li></ul></li><li>图像模型<ul style="list-style-type: none"><li>模型结果可视化，输出图像中与结果相关的区域</li></ul></li><li>大模型<ul style="list-style-type: none"><li>支持自定义情感类型识别</li><li>支持自定义动物类型识别</li><li>能够产生有意义的识别结果置信度</li></ul></li></ul>

指标定义

文本模型、图像模型

指标名	含义	计算公式
精确率 (Precision)	用于衡量模型判定为正类 (违规) 样本中，真正为正类的比例。	$\text{Precision} = \frac{TP}{TP + FP}$
召回率 (Recall)	用于衡量所有实际正类样本中，模型成功识别出的比例。	$\text{Recall} = \frac{TP}{TP + FN}$
F1 分数 (F1 Score)	精确率与召回率的调和平均数，用于在两者之间取得平衡。	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

评测任务提交说明

1. 提交格式：参照示范代码，以邮件形式发送至指定邮箱。
2. 提交次数：每队可提交3次。
3. 结果查看方式：可在指定飞书表格中查看。

## 四、技术参考

- 框架相关
  - PyTorch
    - <https://pytorch.org>
  - Transformers
    - <https://huggingface.co/docs/transformers/index>
- 任务相关
  - 文本分类
    - <https://zhuanlan.zhihu.com/p/598591935>
  - 图像分类
    - <https://zhuanlan.zhihu.com/p/635113065>
- 大模型相关
  - 扣子（大模型） <https://www.coze.cn/open/docs/guides/welcome>