



华为云深度学习在文本分类中的实践

华为 Cloud&AI
李明磊

极客邦科技 会议推荐2019

5月

QCon 北京

全球软件开发大会

大会：5月6-8日
培训：5月9-10日

QCon 广州

全球软件开发大会

培训：5月25-26日
大会：5月27-28日

6月

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE

上海

技术领导力峰会

时间：6月14-15日

GMTC 北京

全球大前端技术大会

大会：6月20-21日
培训：6月22-23日

7月

ArchSummit 深圳

全球架构师峰会

大会：7月12-13日
培训：7月14-15日

10月

QCon 上海

全球软件开发大会

大会：10月17-19日
培训：10月20-21日

11月

GMTC 深圳

全球大前端技术大会

大会：11月8-9日
培训：11月10-11日

AiCon 北京

全球人工智能与机器学习大会

大会：11月21-22日
培训：11月23-24日

12月

ArchSummit 北京

全球架构师峰会

大会：12月6-7日
培训：12月8-9日

目录

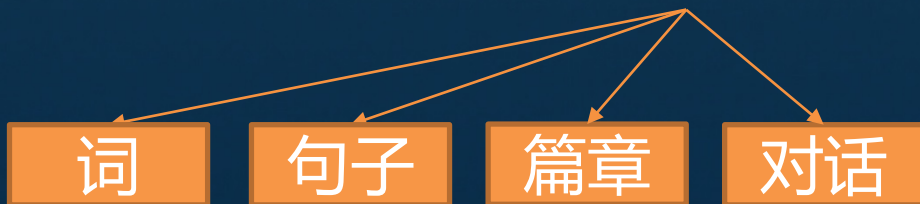


文本分类介绍

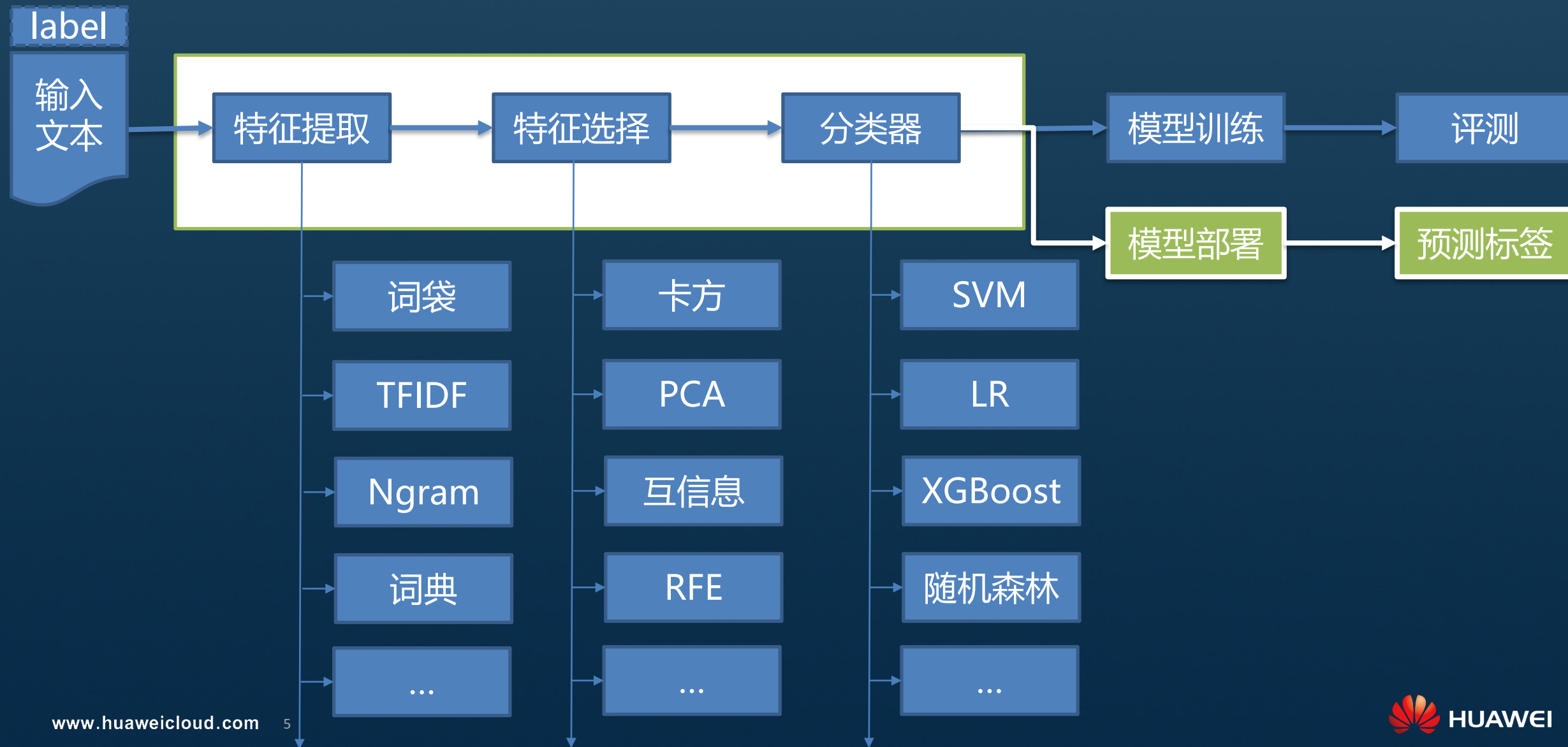
□内容:

- 买没几天就降价一点都不开心, 闪存跑分就五百多点点 --- 😞
- 外观漂亮音质不错, 现在电子产品基本上都是华为的了 --- 😄
- 汽车不错, 省油, 性价比高 --- 😐
- 这个政策好啊, 利国利民 --- 😐
- 电子税务局无法登陆, 提示404。 --- 税务局相关
- 个人所得税APP, 注册的时候操作错误, 怎么办? --- 个税app相关

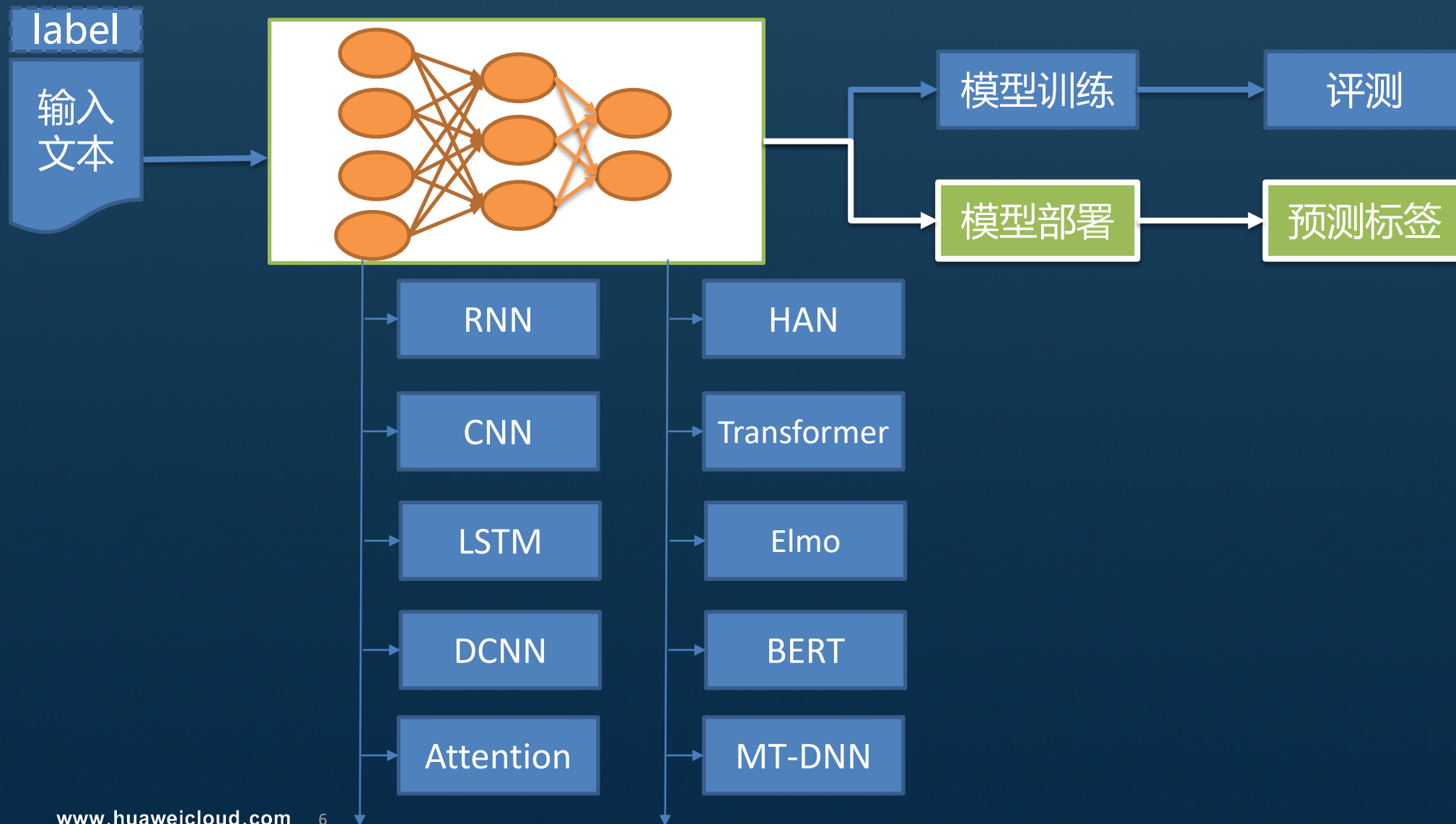
$$f(\text{text}) = \text{label}$$



文本分类方法简史-机器学习

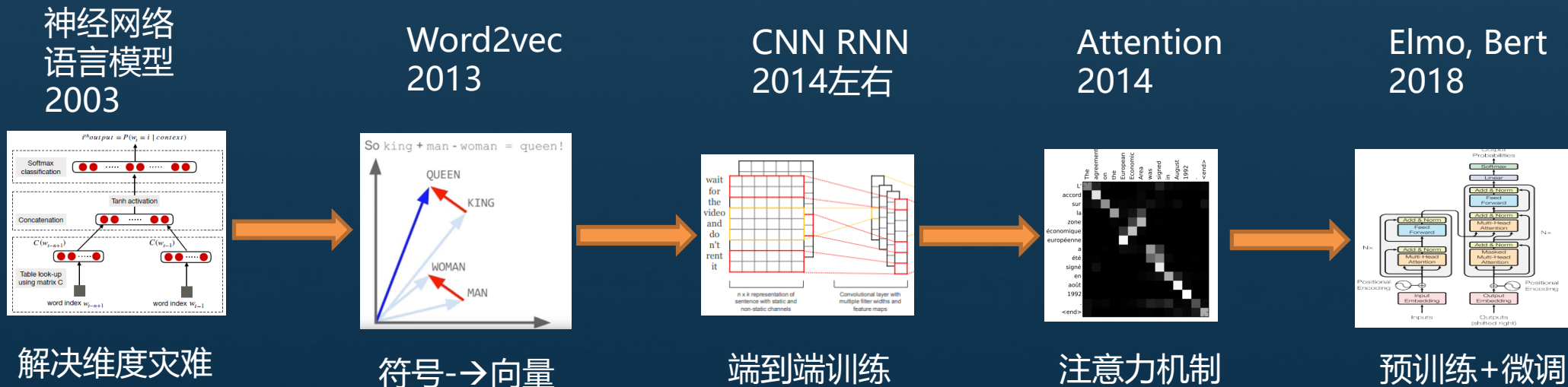


文本分类方法简史-深度学习



文本分类方法简史-深度学习

□ 神经网络NLP里程碑：



预训练+微调

大规模语料训练通用语言模型

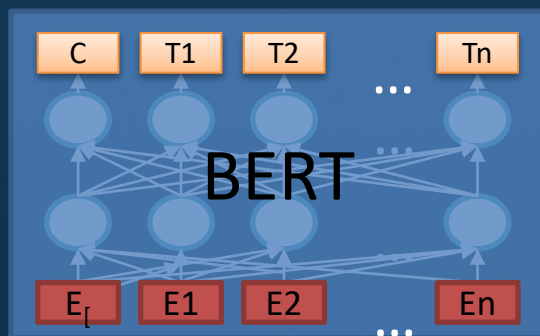


在目标语料上微调语言模型



在目标语料上训练分类器

模型：



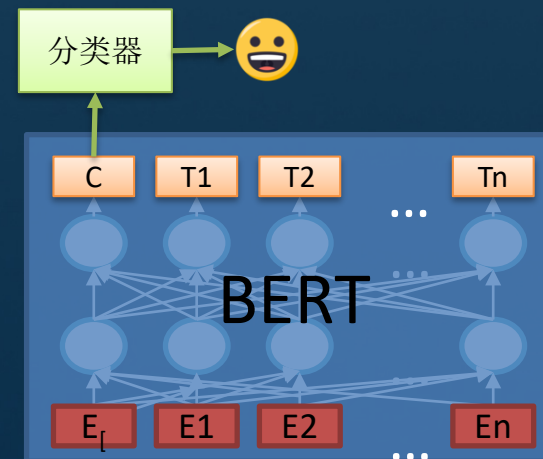
数据：



目标：

预测mask词和下一句

模型：



数据：

手机不错，高大上	正面
手机太差劲了，又贵又卡	负面
续航给力，价格实在	正面

目录

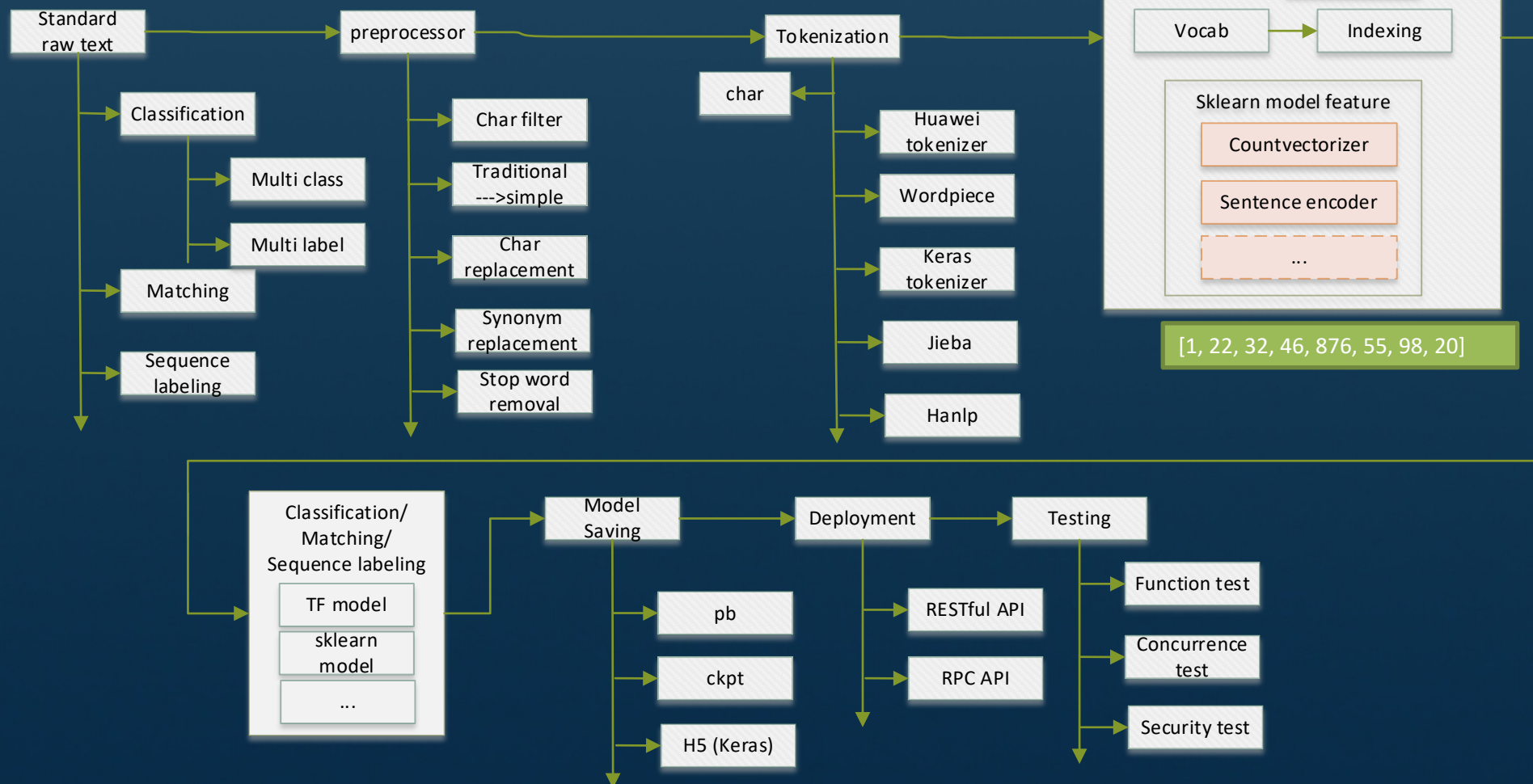


深度学习框架

手機不錯，高大上

手机不错，高大上

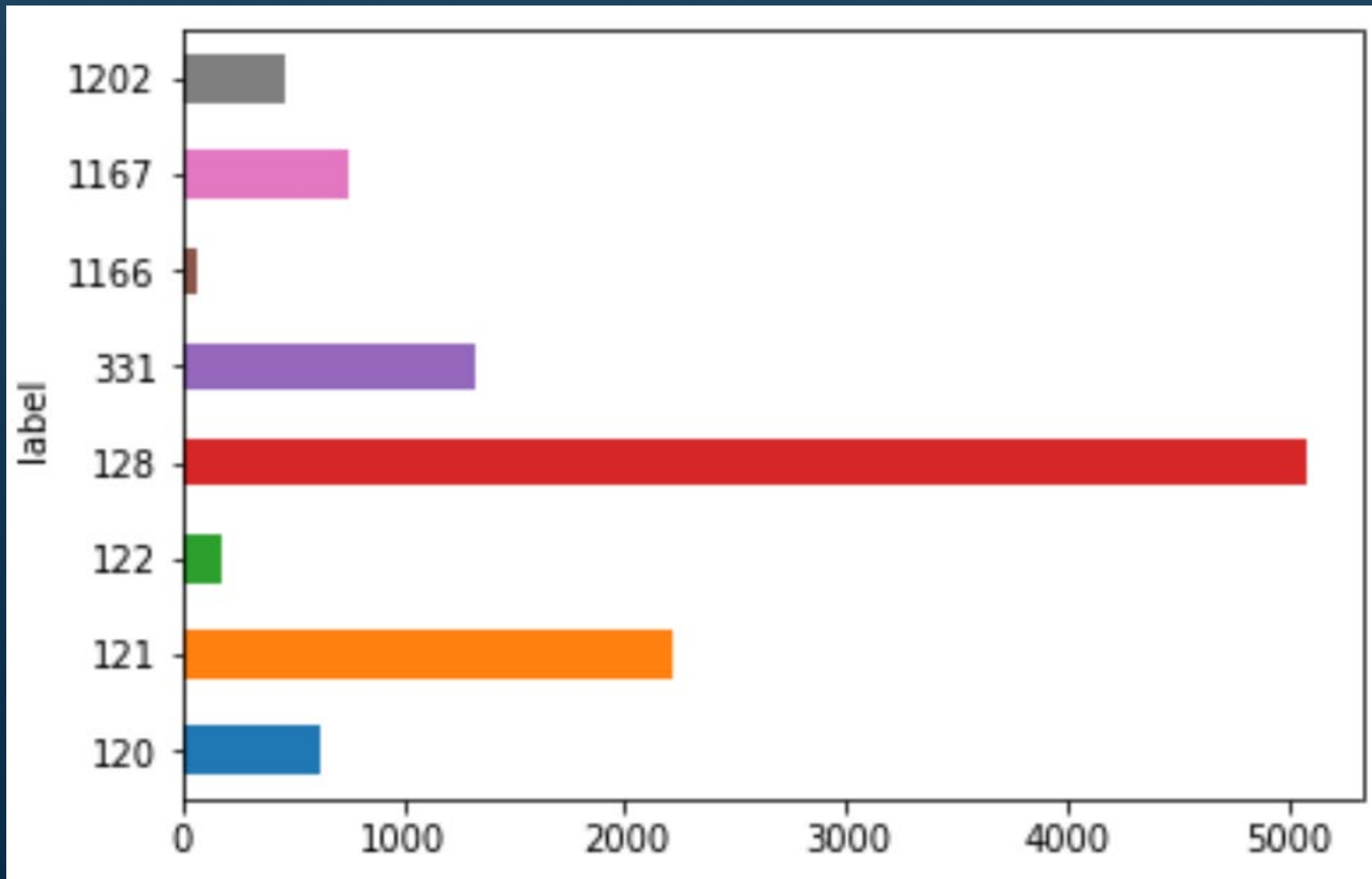
[手机不错，高大上]



目录



数据不均衡



数据不均衡

□ 预处理方法

- 上采样
- 下采样
- SMOTE
- 数据增广

□ 集成方法

- SMOTEbagging

□ 改损失函数

- Focal loss

“An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.” *Information Sciences* 250 (November 20, 2013): 113–41.

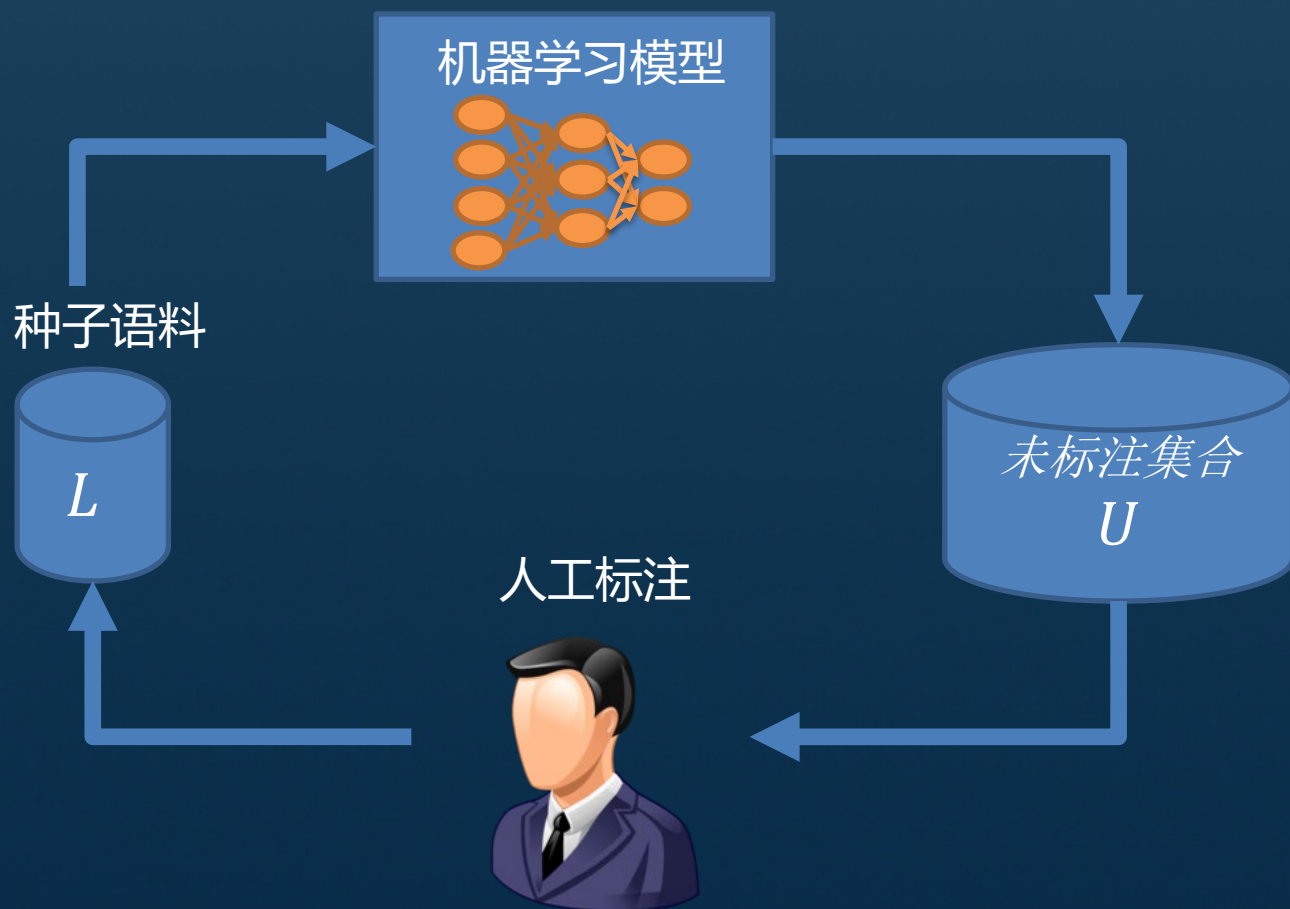
$$\text{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t).$$

- 更多数据：首先，在当前任何实际环境中正则化模型的最好方式是增加更多真实的训练数据。在你能收集更多数据时，花费大量工程时间试图从小数据集上取得更好结果是很常见的一个错误。我认为增加更多数据是单调提升一个较好配置神经网络性能的唯一可靠方式。

----特斯拉人工智能主管Andrej Karpathy

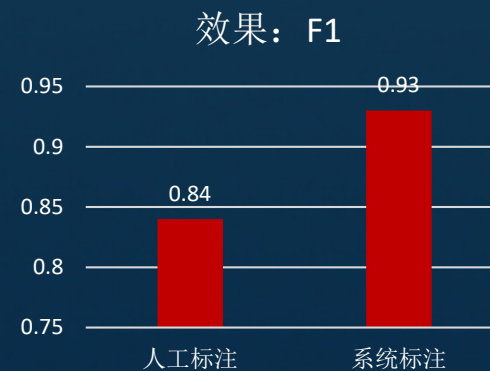
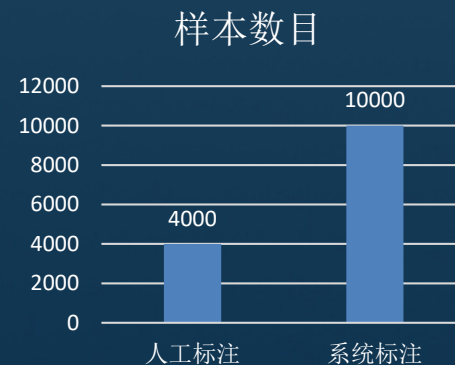
数据标注成本高

□ 主动学习框架：




□ 选择策略：

➢ 基于置信度



华为云主动学习平台

 HUAWEI

香港

控制台 服务列表 收藏

Q 费用 资源 工单 企业 备案 User Name

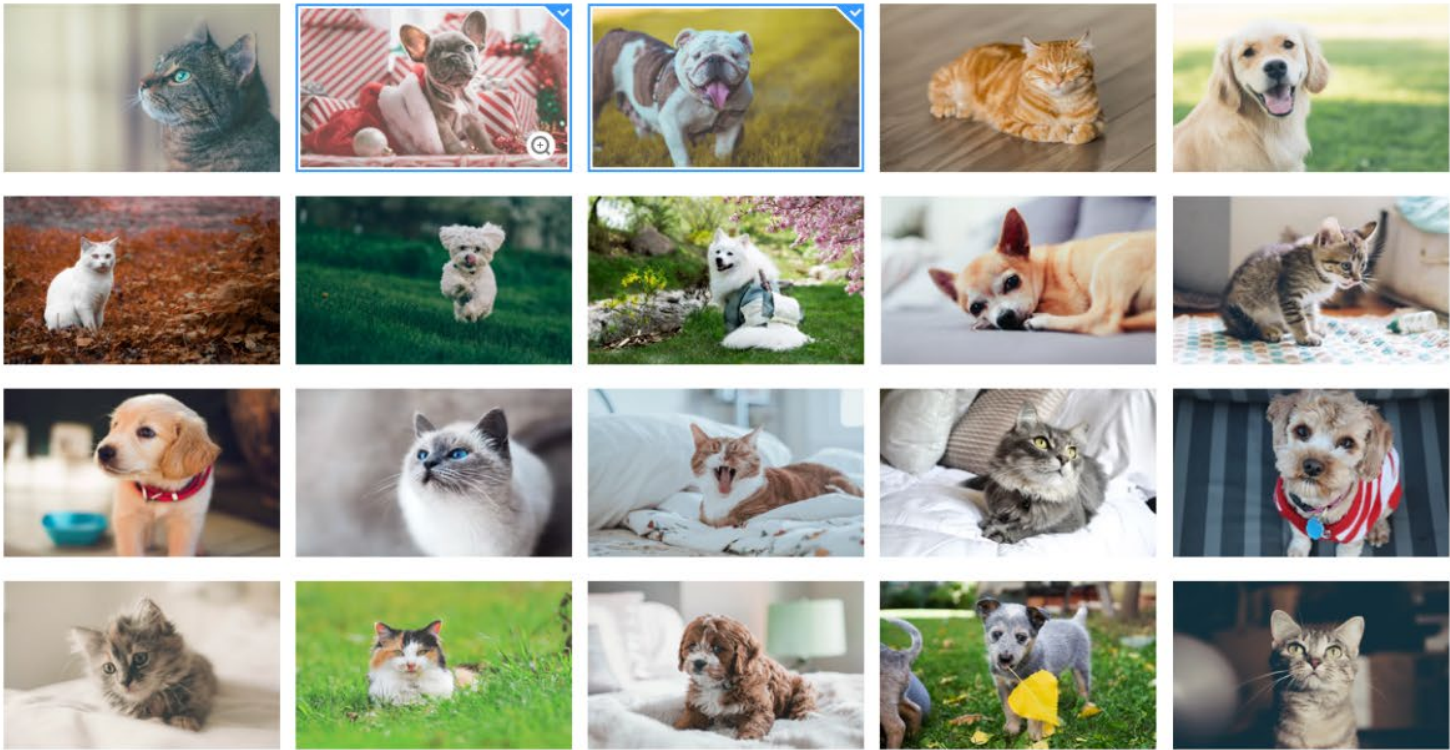
Cat and Dog < 返回数据标注

启动智能标注 保存数据集

已标注 58 未标注 153

添加图片 删除图片 同步数据源

☐ 选择所有



25 总条数: 153 < 1 2 3 4 5 ... 16 >

添加标签

C 请输入标签 添加


快捷键 标签

选中文件标签 当前中 2 张图片

Dog 2 C

将选中图片确认为已标注

华为云主动学习平台

 HUAWEI

香港

控制台

服务列表

收藏

Q

费用

资源

工单

企业

备案

User Name

1

Cat and Dog

< 返回数据标注

启动智能标注


保存数据集

已标注 58

未确认 153

智能标注进度: 86%(86/100)

停止



正在推理, 请稍后
耗时3分30秒.....

添加标签

C

请输入标签

添加

快捷键 标签

选中文件标签

当前中 0 张图片

将选中图片确认为已标注

目录



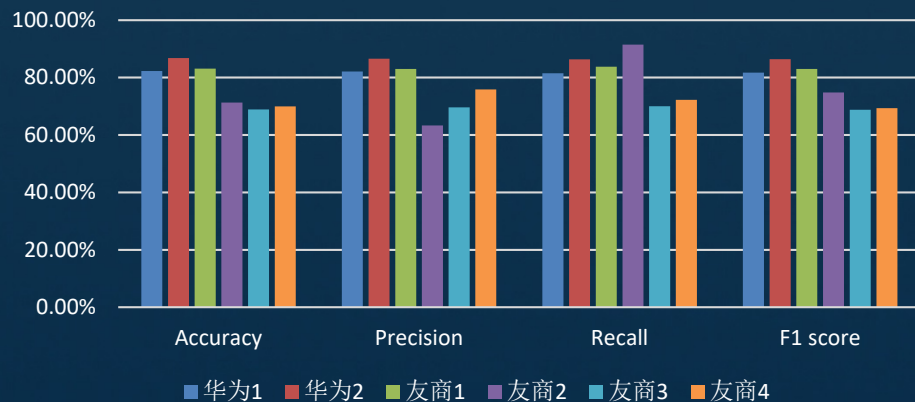
情感分析

□ 内容:

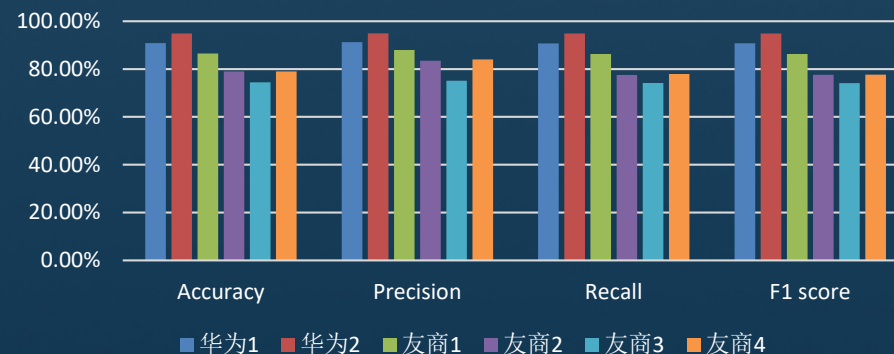
- 买没几天就降价一点都不开心，闪存跑分就五百多点点 --- 😞
- 外观漂亮音质不错，现在电子产品基本上都是华为的了 --- 😊
- 汽车不错，省油，性价比高 --- 😊
- 这个政策好啊，利国利民 --- 😊



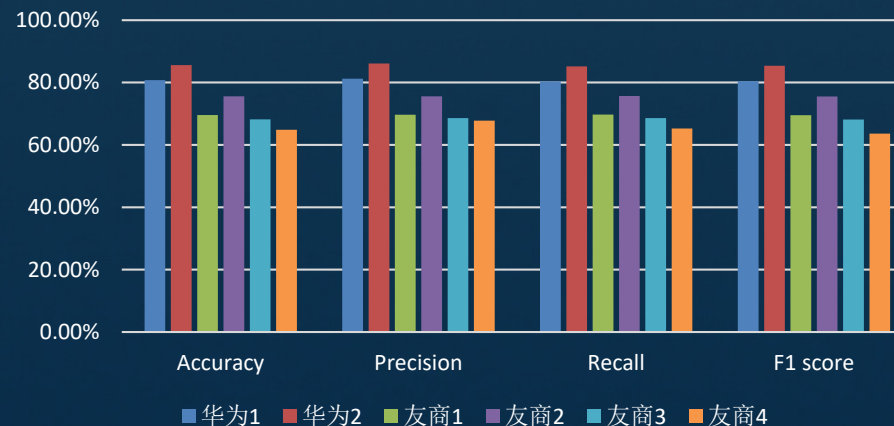
和友商效果对比-汽车领域



和友商效果对比-电商领域



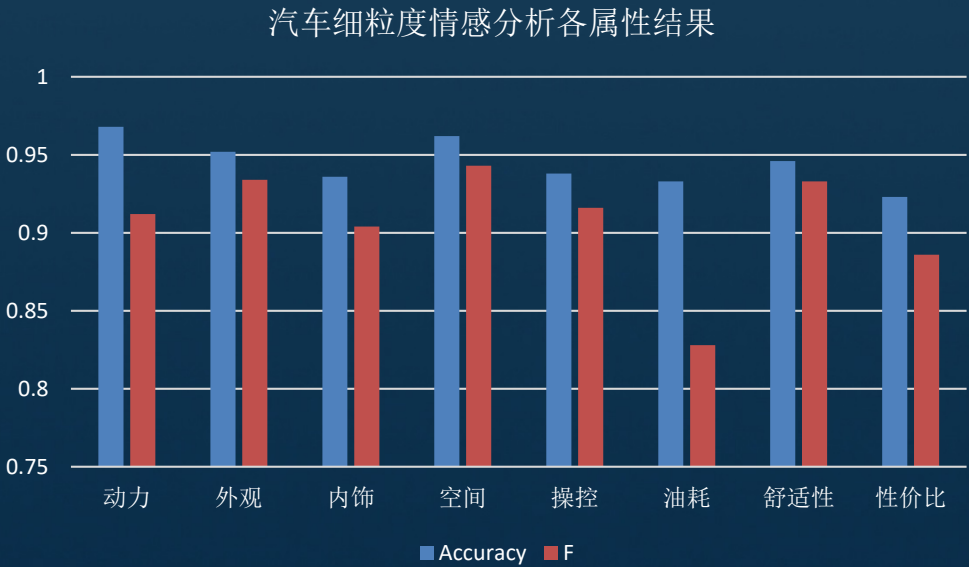
和友商效果对比-社交领域



细粒度情感分析

评论	动力	外观	空间	油耗
2.0T涡轮增压发动机动力强，高速120超车没压力；外观是我和老婆都比较喜欢的款；后排空间有点小；有点费油啊。	😄	😄	😞	😞

- ❑ 定制化Loss，单模型多输出
- ❑ 数据标注灵活
- ❑ 结合数据增强，针对不均衡数据做优化



其他分类案例

广告检测



识别文本是否是广告。如“去屑洗发水，全国包邮”。

准确率：92%

税务问题分类



识别用户在税务局中咨询的问题类型，并进行热点问题分析。

准确率：99%

客服话题分类



识别客户对话过程用用户反馈的话题类型，并进行热点话题分析等。

准确率：96%

案件描述分类



对案件描述进行分类，并进行可视化展示。

准确率：93%

政务问题分类



识别用户所问问题类型并进行热点问题分析。

准确率：98%

EI体验空间





全球技术领导力峰会

Geekbang> | TGO 鲲鹏会
极客邦科技

500+ 高端科技领导者与你一起探讨 技术、管理与商业那些事儿



🕒 2019年6月14-15日 | 📍 上海圣诺亚皇冠假日酒店



扫码了解更多信息



Thank You.

Copyright©2018 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.