QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2018

《互联网文本内容安全：一种对抗式AI设计实践》

演讲者 / 王国印[腾讯云-专家级研究员]

主办方 Geekbang. 极客邦科技 InfoQ

个人介绍

2007.9-2009.6
CGCL

2013.9-2015.5
AI+电商
阿里巴巴
全球购业务
标签导购业务
闲鱼业务

2015.5-2017.10
AI+物流
菜鸟网络
地址库业务
菜鸟裹裹业务
菜鸟驿站业务
自提柜业务

2018.1-至今
AI+业务安全
腾讯云
电商防黄牛
内容安全

腾讯云-天御 专家研究员

提纲

1. 腾讯云-天御
2. 内容安全现状
3. 带来的影响
4. 天御解决方案
5. 思考&总结

腾讯云-天御：AI+业务安全

金融风控

金融（银行、互金、保险等）

虚假身份、撸口子

身份核验、借贷反欺诈

内容安全

视频（直播）

互联网（社区）

广告、涉黄、涉政、违法

图像安全鉴定、反垃圾

流量风控

互联网（电商、O2O）

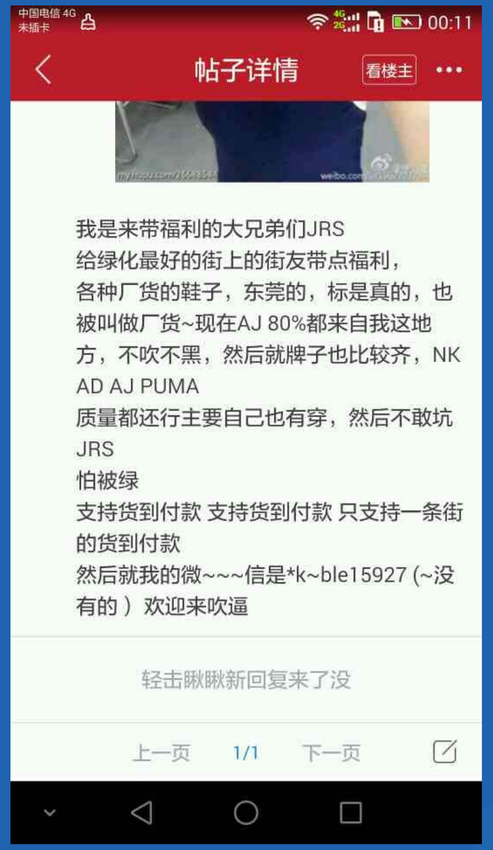
传统行业（快消、新零售）

金融（银行、互金）

羊毛党、黄牛党

防刷、流量保护、验证码

内容安全现状：各类社区被违规内容充斥



违规平台被媒体曝光

央视曝光涉黄直播平台勾当

2017-04-17 14:45:55 来源: 北京晚报(北京)

分享到：

（原标题：央视曝光涉黄直播平台勾当）

去年开始推送直播秀链接。 视频截图

网络低俗直播、色情直播，这些突破道德底线与法律红线的网络直播，污染了网络空间。国家相关部门一直对网络低俗色情直播持续整治打击，但在利益诱惑之下，乱象屡禁不止。央视记者经过三个多月的跟踪调查，发现某些知名网站、用户影响力巨大的社交媒体成为了低俗内容的传播工具；某些新闻客户端、自媒体直播平台竟成为低俗直播的入口，频频向用户推送。

“净网”行动破获非法广告联盟传播淫秽物品案

张 贺 姚祎婷

《人民日报》（2017年08月30日 18版）

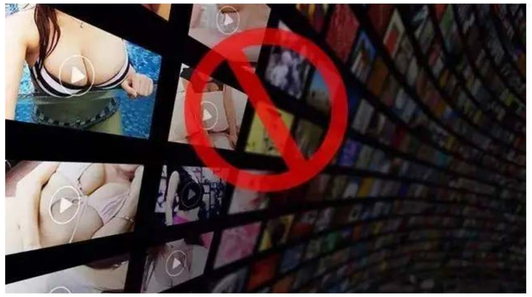
近期，在“净网2017”专项行动中，由公安部和全国“扫黄打非”办公室联合督办的江苏宿迁“2·07”王某传播淫秽物品牟利案告破，一条由淫秽网站与非法广告联盟相互勾结，大肆传播淫秽、色情以及贩卖公民个人隐私信息等内容的利益链浮出水面。“2·07”专案组历经3个多月的侦查，转战6省7市和澳门特别行政区，抓获犯罪嫌疑人50人，成功摧毁淫秽网站站点118处，涉案总金额超2000万元。

据民警介绍，今年1月，宿迁市公安局治安支队接全国“扫黄打非”办交办“78美术网”涉嫌传播淫秽色情及低俗信息案件线索后，开展核查取证工作，并于2月7日抓获犯罪嫌疑人王某。为了解王某开设淫秽网站的目的，专案组对网站页面大量广告信息进行专题分析，结合对王某的突击审讯，发现一条“广告商—互联网广告联盟—网站主网站发布”的犯罪利益链，与王某合作的“富投联盟”等4个广告联盟浮出水面。据悉，专案组还彻查了淫秽图片来源，抓获了犯罪嫌疑人莫某等。

走了黄鳝，又掏黄瓜！涉黄直播屡禁不止，直播平台路在何方？

传媒1号 2017-05-29 关注作者 原文地址 转藏本文

来源 | 新榜（ID:newrankcn）



继广电总局、网信办、“扫黄打非”办后，这次轮到文化部对直播平台下手了。《文化部关停10家直播平台，“花椒”虚假故宫直播被处罚》——近日，大量行业媒体转载了人民日报客户端这则报道：

涉政、涉黄触及国家红线

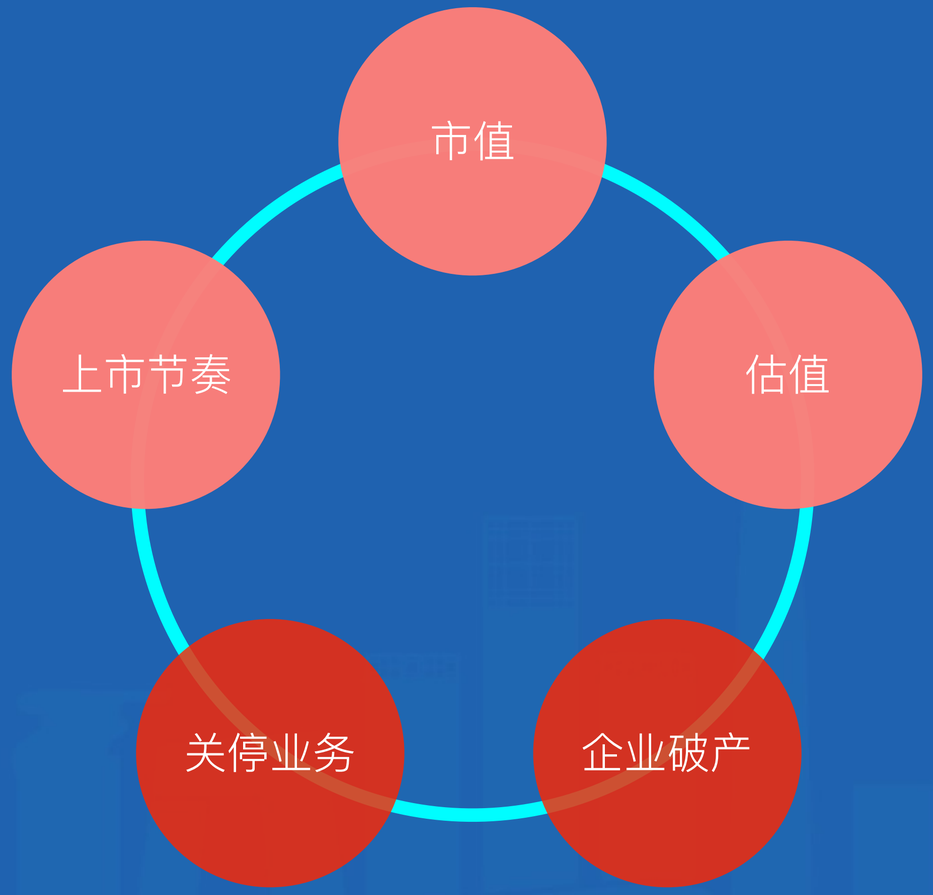
常见管控方式

- **约谈**
- **限期整改**
- **下架APP**
- **永久关停**

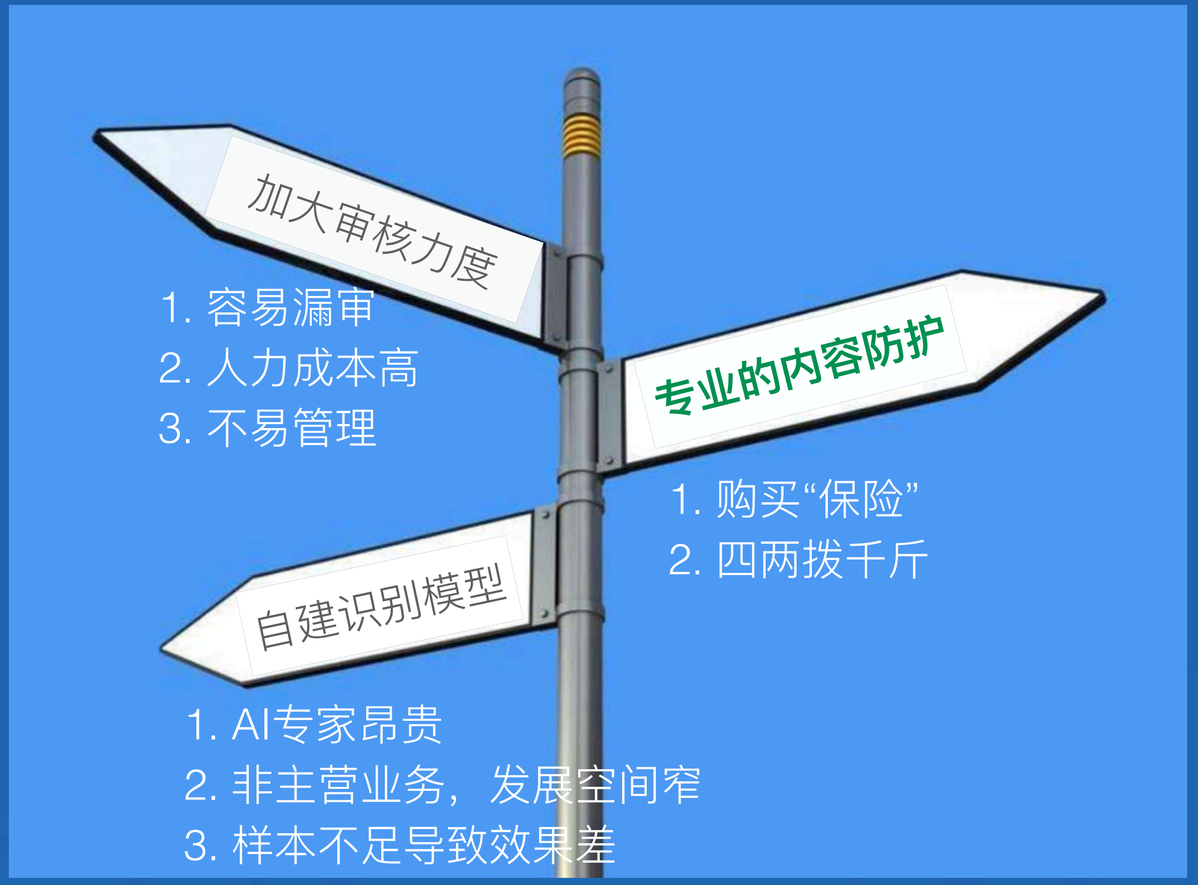
网信办2017战果

- 约谈**2003**家企业
- 关闭**22,587**家网站
- 移送**2045**件
- 关闭**317万**个网络账号

内容安全对企业的影响



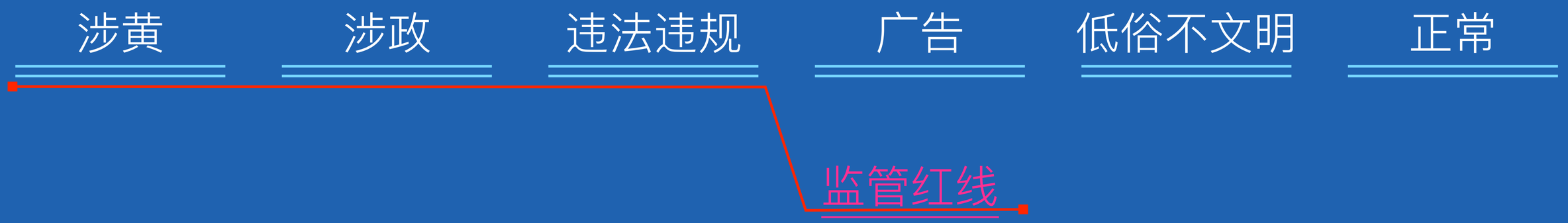
面对威胁企业如何抉择


加大审核力度
1. 容易漏审
2. 人力成本高
3. 不易管理
专业的内容防护
1. 购买“保险”
2. 四两拨千斤
自建识别模型
1. AI专家昂贵
2. 非主营业务，发展空间窄
3. 样本不足导致效果差

现有解决方案

关键词过滤	文法过滤	静态机器学习模型	混合策略
歧义导致误杀	上下文不易枚举	模型很容易失效	应对不了变种问题
无法利用沉淀数据	靠人工归纳，持续投入		
新客户接入成本高	未登录case处理弱		

UGC分类



UGC特点


规模巨大
对抗性
变种绕过防控策略
防控尺度
因业务类型而异
腾讯云天御应对策略
高并发算法引擎
弹性部署
降低人力
审核人力、研发投入
数据闭环
T+1更新策略
分布式存储
模型个性化

系统架构

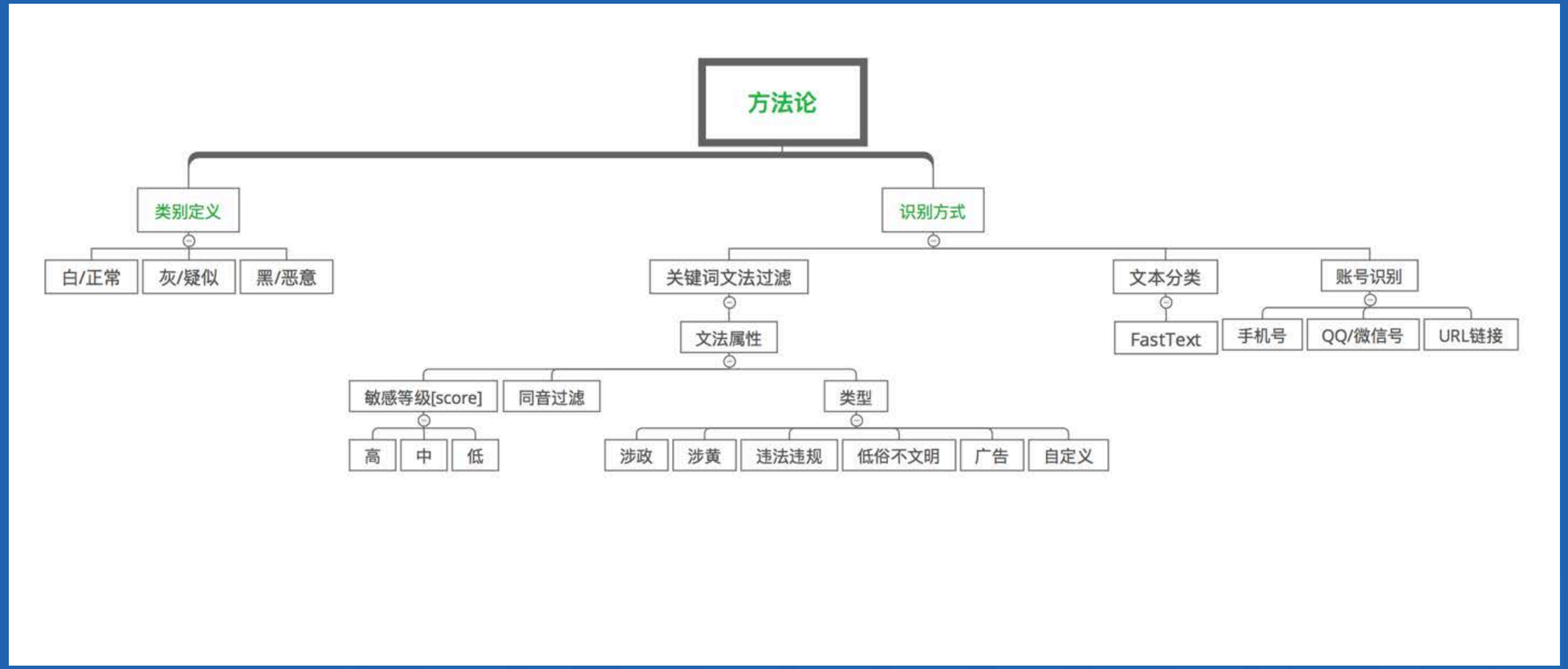
支持自更新

模型自动升级

新旧模型秒级替换

支持多模型训练实现个性化定制

垃圾识别


方法论
类别定义
白/正常
灰/疑似
黑/恶意
识别方式
关键词文法过滤
文法属性
敏感等级[score]
高
中
低
同音过滤
类型
涉政
涉黄
违法违规
低俗不文明
广告
自定义
文本分类
FastText
账号识别
手机号
QQ/微信号
URL链接

文本分类算法选型： FastText

- Embedding
- N-gram Feature
- Simple
- Fast
- Nice Performance

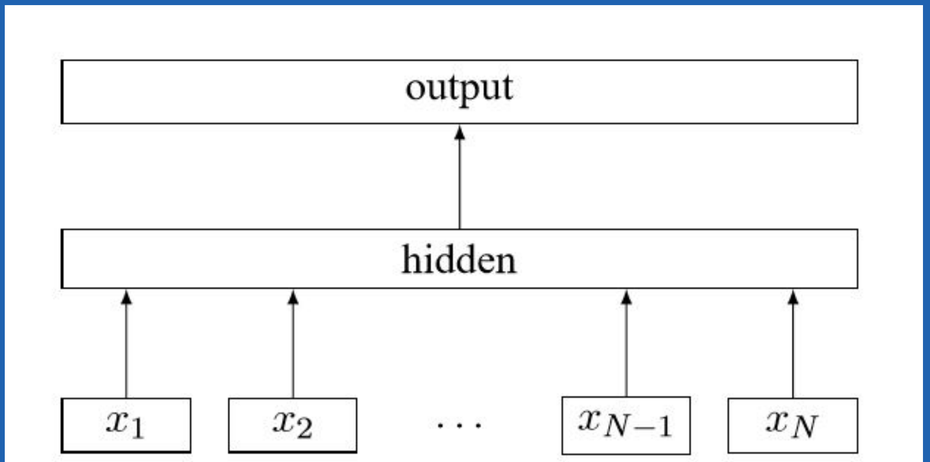


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

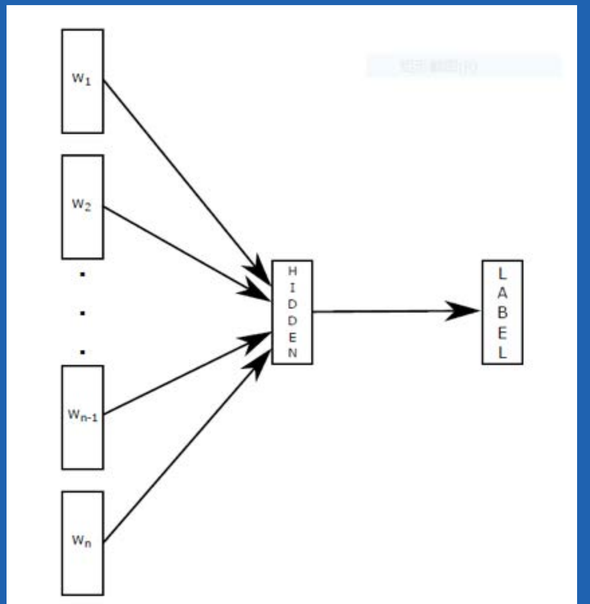


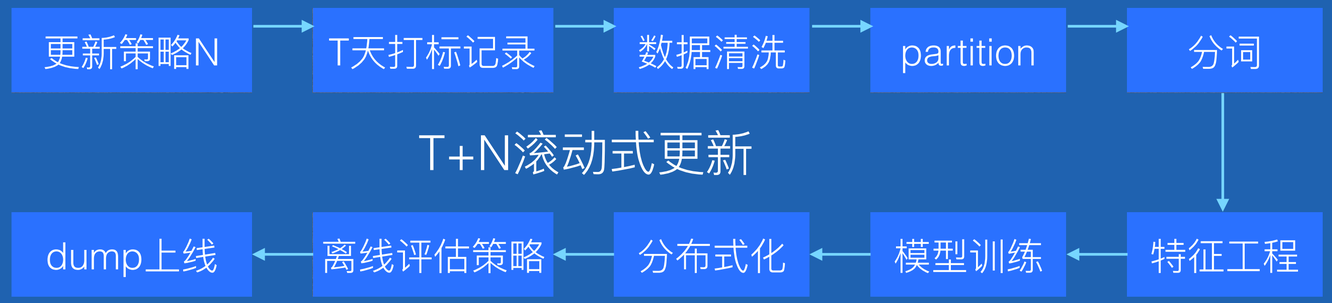
Figure 1: Model architecture for fast sentence classification.

Performance

(mulit-label label prediction task,ask to prediction top5, 3 million training data,full score:0.5)

Model	fastText	TextCNN	TextRNN	RCNN	HierAtteNet	Seq2seqAttn	EntityNet	DynamicMemory	Transformer
Score	0.362	0.405	0.358	0.395	0.398	0.322	0.400	0.392	0.322
Training	10m	2h	10h	2h	2h	3h	3h	5h	7h

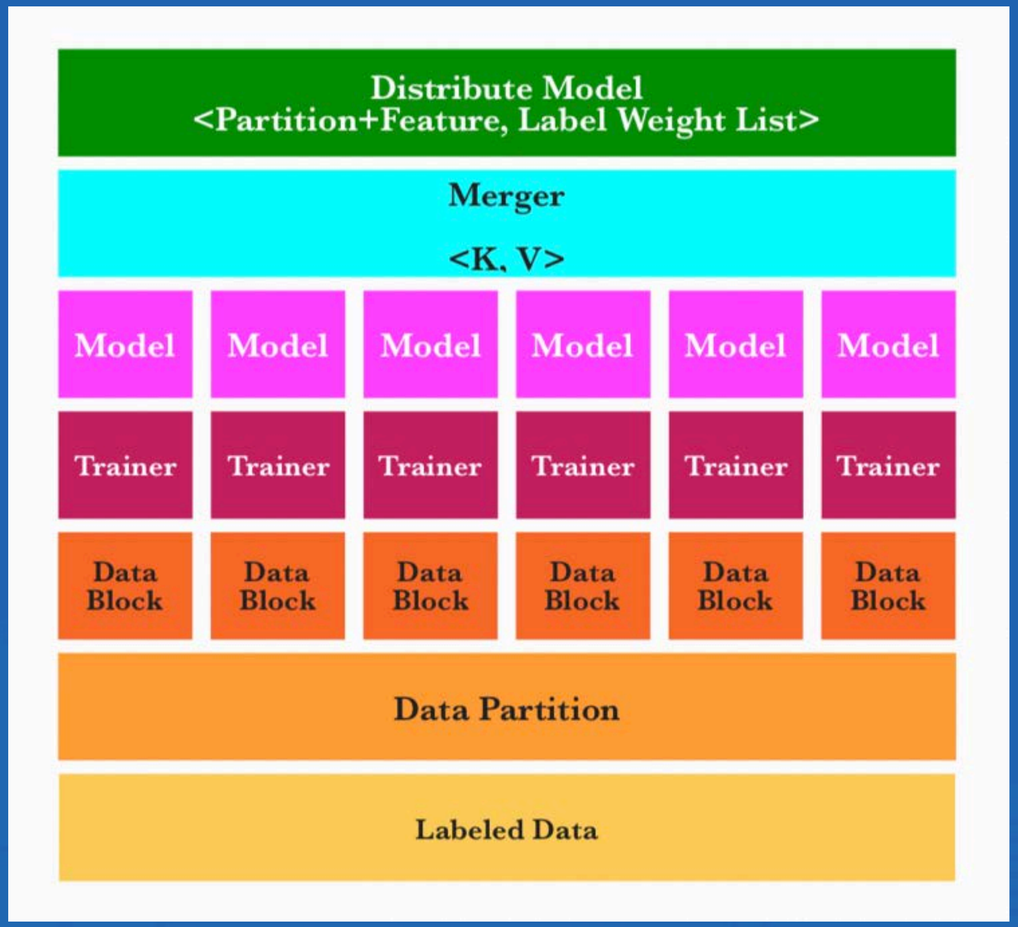
模型个性化



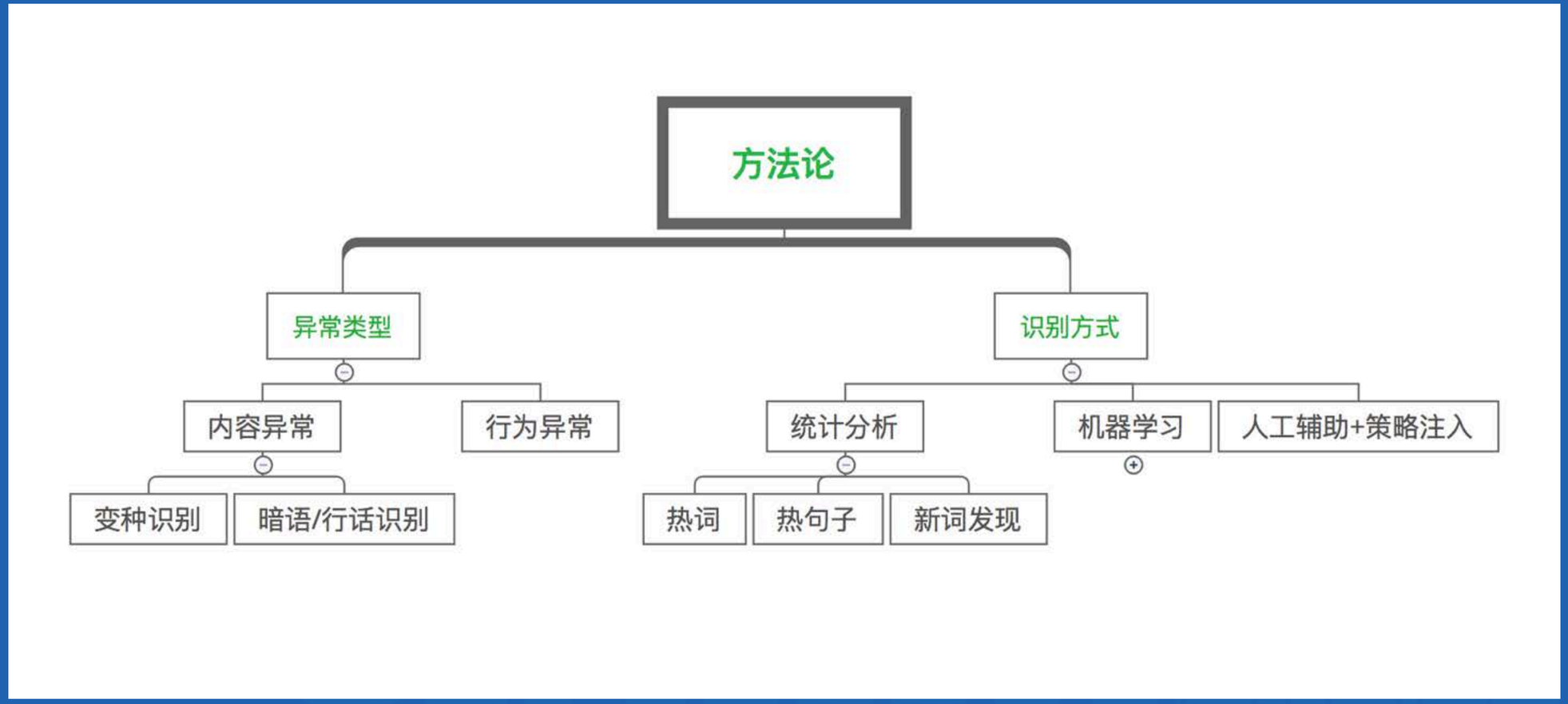
模型自动训练　　模型随对抗而升级

模型自动更新　　分布式存储实现秒级更新GB级模型

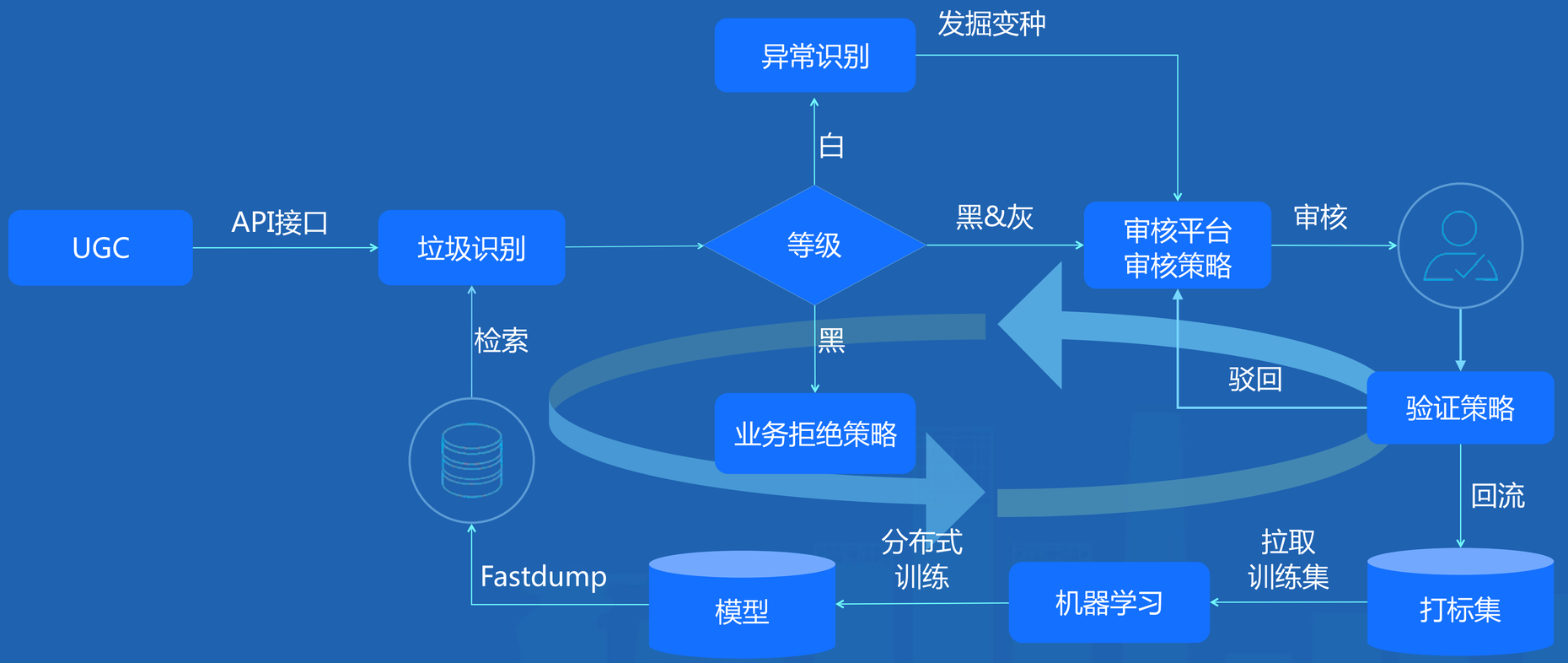
大大降低研发投入



异常识别


方法论
异常类型
识别方式
内容异常
行为异常
统计分析
机器学习
人工辅助+策略注入
变种识别
暗语/行话识别
热词
热句子
新词发现

数据闭环



数据闭环为模型滚动式更新创造条件

审核抽样策略降低人工审核量

系统指标

封闭测试准确率

封闭测试召回率

封闭测试准确度

准确率

召回率

抽样准确率

进审量(条/天)

盲审抽样率

盲审一致率

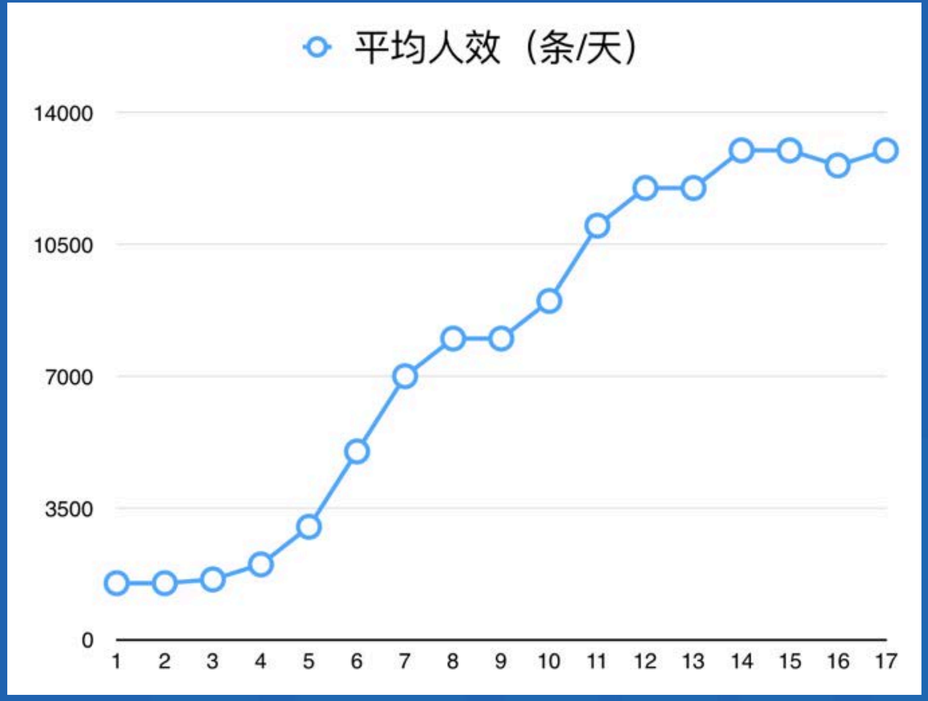
人效（条/小时）

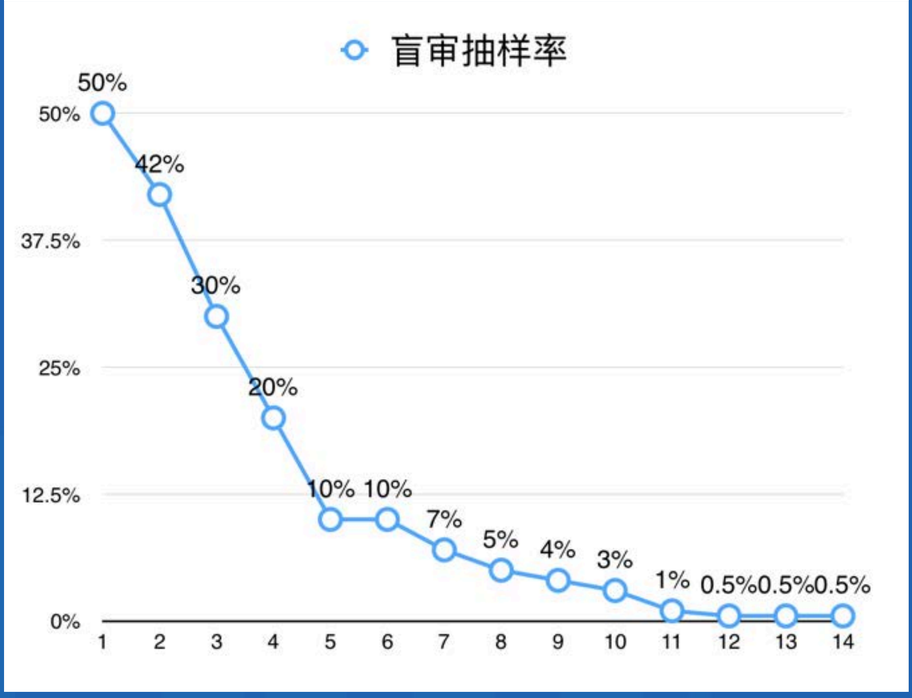
审核平均延时

总结

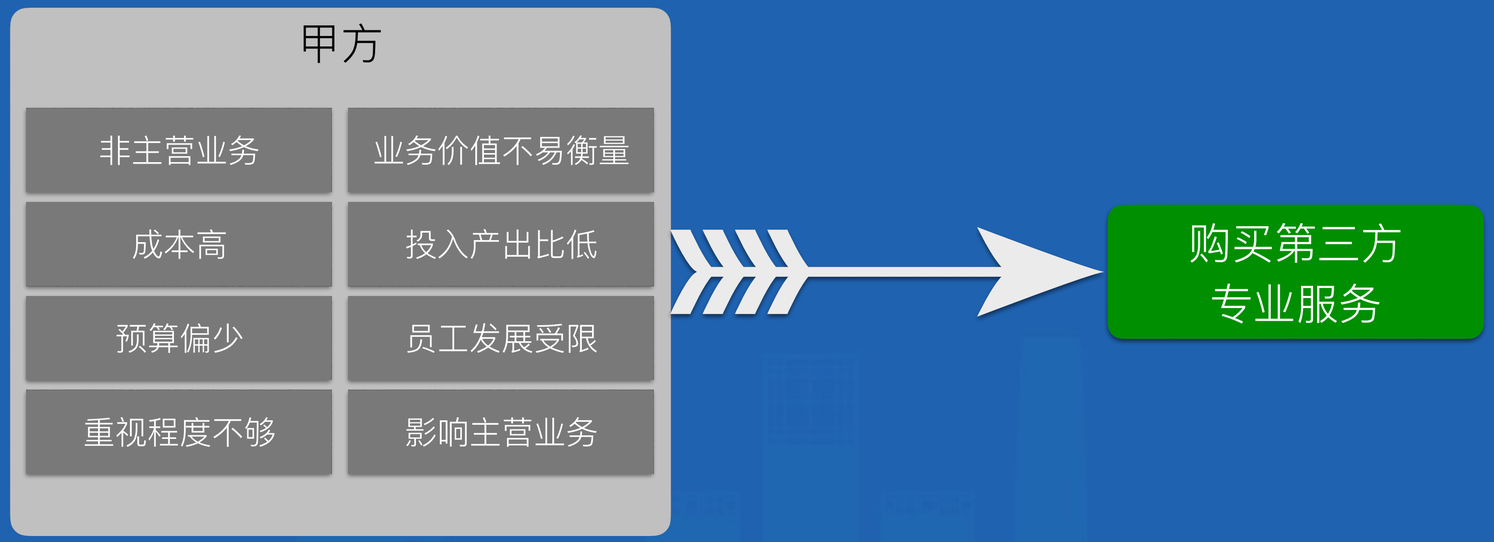
抽样准确率

- 涉黄：**99.7**%
- 涉政：**99.9**%
- 违法违规：**99.9**%
- 广告：**97.3**%
- 低俗不文明：**99**%





思考


甲方
非主营业务
业务价值不易衡量
成本高
投入产出比低
预算偏少
员工发展受限
重视程度不够
影响主营业务
购买第三方
专业服务