河海大学
HOHAI UNIVERSITY

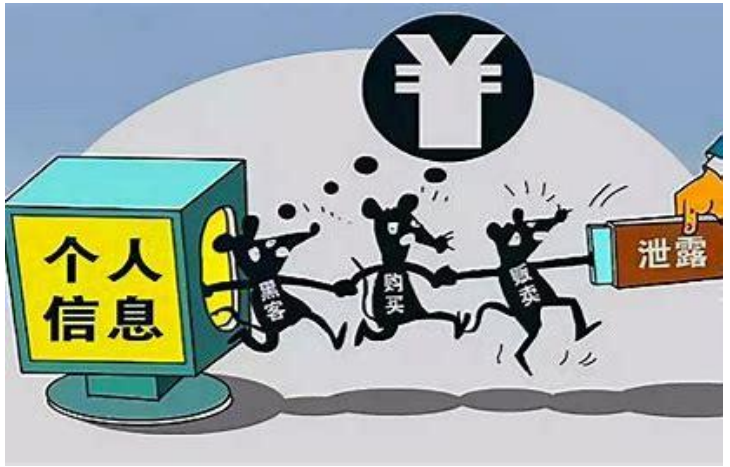
基于LDA的社交应用隐私政策合规性评价研究

汇报人：徐绪堪 2023年7月12日

河海大学商学院
河海大学统计与数据科学研究所
常州市工业大数据与知识管理重点实验室

目录 CONTENTS

2019年12月，工信部发布了第一批**侵害用户权益**的社交应用名单。该名单包含41款社交应用程序。2020年7月，北京互联网法院对抖音和微信读书两款社交应用进行了一审判决，认定其存在侵害用户个人信息的行为，次年12月，**工信部下架了106款侵害用户权益的社交应用。**



App个人信息泄露事件频发，引起国家的高度重视！

国家网信办：下架“滴滴企业版” “滴滴顺风车”等25款App

2021年07月09日22:20 | 来源：人民网 小字号


中华人民共和国中央人民政府 www.gov.cn
首页 | 繁体 | 英文EN | 登录 | 邮箱

关于下架侵害用户权益APP名单的通报

2021-12-12 16:47 来源：工业和信息化部网站 字号：默认 大 超大 打印

今年以来，工业和信息化部持续推进APP侵害用户权益专项整治行动，加大常态化检查力度，先后三次组织对用户反映强烈的重点问题开展“回头看”。11月3日，工业和信息化部针对APP超范围、高频次索取权限，非服务场景所必需收集用户个人信息，欺骗误导用户下载等违规行为进行了检查，并对未按要求完成整改的APP进行了公开通报。截至目前，尚有5款APP未按工业和信息化部要求完成整改（详见附件1）。各通信管理局按照工业和信息化部统筹部署，积极开展APP技术检测，截至目前尚有101款APP仍未完成整改[截图(Alt + A)]件2-8）。

依据《个人信息保护法》《网络安全法》等相关法律要求，工业和信息化部组织对上述共计106款APP进行下架，相关应用商店应在本通报发布后，立即组织对名单中应用软件进行下架处理。针对部分违规情节严重、拒不整改的APP，属地通信管理局应对APP运营主体依法予以行政处罚。


中华人民共和国中央人民政府 www.gov.cn
首页 | 繁体 | 英文EN | 登录 | 邮箱

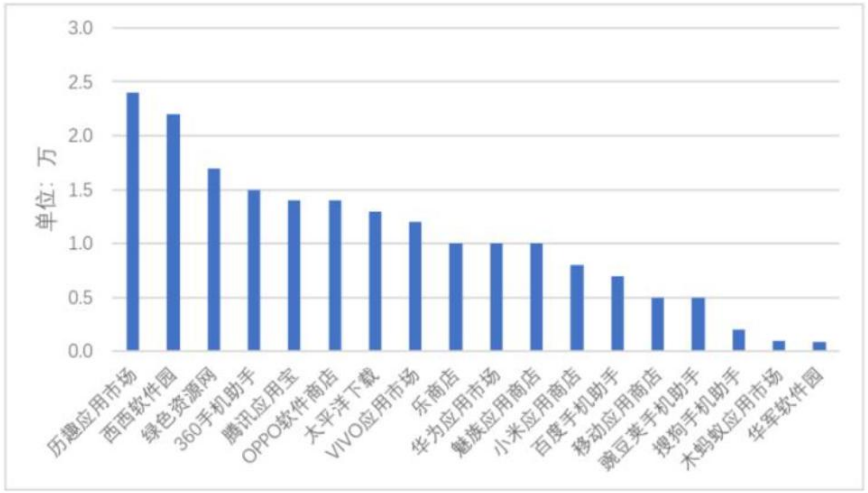
开展侵害用户权益整治行动 工信部下架540款APP

2021-11-18 09:25 来源：人民日报 字号：默认 大 超大 打印

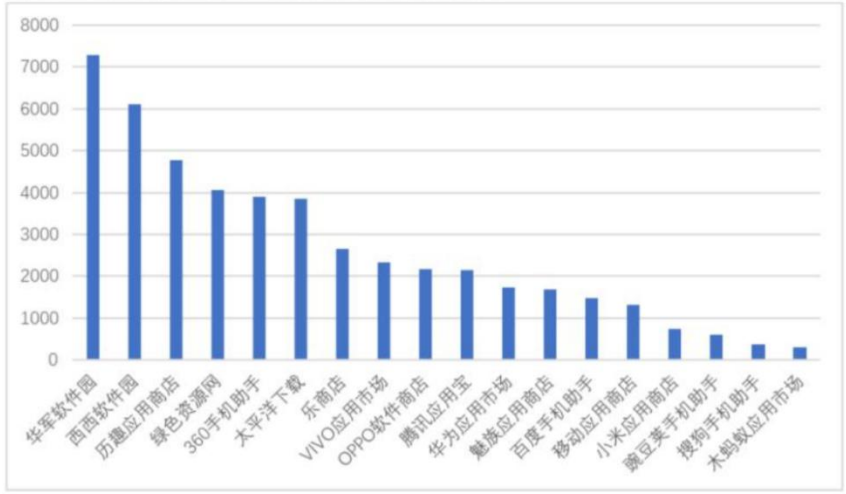
“截至目前，已组织检测21批次共244万款APP，累计通报2049款违规APP，下架540款拒不整改的APP，对违规行为持续保持高压震慑。”工业和信息化部信息通信管理局有关负责人日前透露，工信部聚焦人民群众反映强烈的APP违规处理用户个人信息、设置障碍、骚扰用户、欺骗误导用户等问题，纵深推进APP侵害用户权益整治行动，取得显著成效。

工信部不断强化应用商店关键责任链管理，督促应用商店加强自查清理，应用商店已主动下架40余万款违规APP。据介绍，近期工信部又部署开展了服务感知提升行动，推动全行业优化服务举措、提升服务能力，建立个人信息保护“双清单”，持续加大用户信息保护力度。

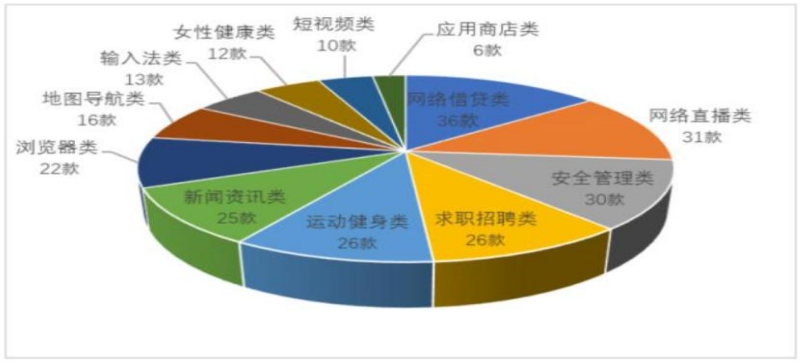
过渡索取权限



强制收集用户信息



超范围收集个人信息



个人隐私泄露风险问题依然严峻

数据来源：《App个人信息泄露情况调查报告》

如何**有效分析隐私问题**，**有效保护用户个人信息**并帮助平台建设隐私服务已经成为当前的研究热点。

目前，个人隐私政策保护问题的相关研究可分为三个方面，**个人隐私保护制度设计研究，隐私政策的内容分析研究，隐私政策的实际应用研究**。个人隐私保护制度设计研究为内容分析设计提供了实践基础。隐私政策内容研究将相关的理论、方法和技术应用结合起来，以引导相关理论和技术研究。这三个方面的研究相互促进，构成了一个完整的研究体系。

1.研究对象

将不同的社交应用分为即时交流、博客论坛、婚恋交友、购物生活、职场社交、视频社区和兴趣社交七个类别，经过对社交App进行筛选，最终确定研究对象为28个手机App

对隐私政策、未成年人保护政策和第三方SDK库文本进行整理，最终得到共49份隐私政策相关文本。

社交应用类型	名称	社交应用类型	名称
即时交流	微信	职场社交	领英职场
	QQ		BOSS直聘
博客论坛	钉钉	视频社区	脉脉
	微博		抖音
	知乎		爱奇艺
	百度贴吧		腾讯视频
婚恋交友	小红书	兴趣社交	快手
	世纪佳缘		斗鱼
	珍爱		虎牙
	Soul		虎扑
	探探		豆瓣
	MOMO陌陌		Keep
购物生活	淘宝		酷安

1.研究对象

从命名方式来看，大多数隐私政策文本命名均包含隐私和个人信息字样，方便用户快速寻找和定位；从篇幅字数来看，大多数政策文本的文字篇幅都超过了10000字，篇幅最长的是Boss直聘，全文达到了17995字，政策文本的平均篇幅为13765字，标准差为3341.3，可以发现社交应用隐私政策文本的**内容篇幅差别较大**，且随着时间的推移各应用平台也在根据相关法律完善隐私政策，不断提高政策覆盖度与完整性。

2.数据预处理

（1）数据清洗

通过jieba包对文本进行分词，再利用哈工大停用词表对文本进行第一次停用词处理

（2）特征提取

计算词语的Tf-idf值，获得高频词语的统计结果

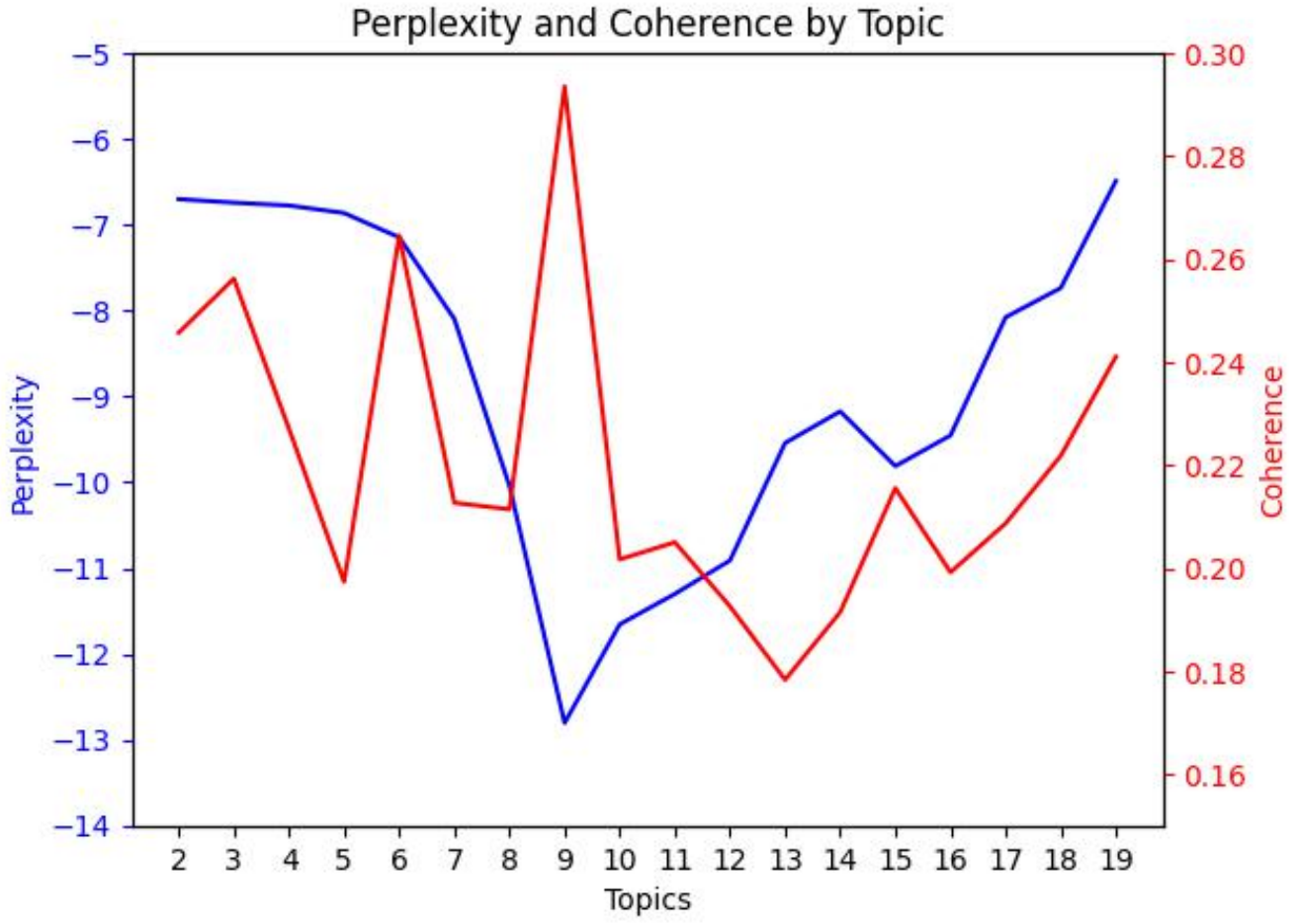
（3）停用词去除

通过高频词语的统计结果可以发现，“信息”，“个人信息”，“政策”等70个词语在文本中出现次数过多而本身对本次研究影响不大，故此步骤采用局部去除这部分高频词，进一步去除停用词，防止其他词语被覆盖的做法。

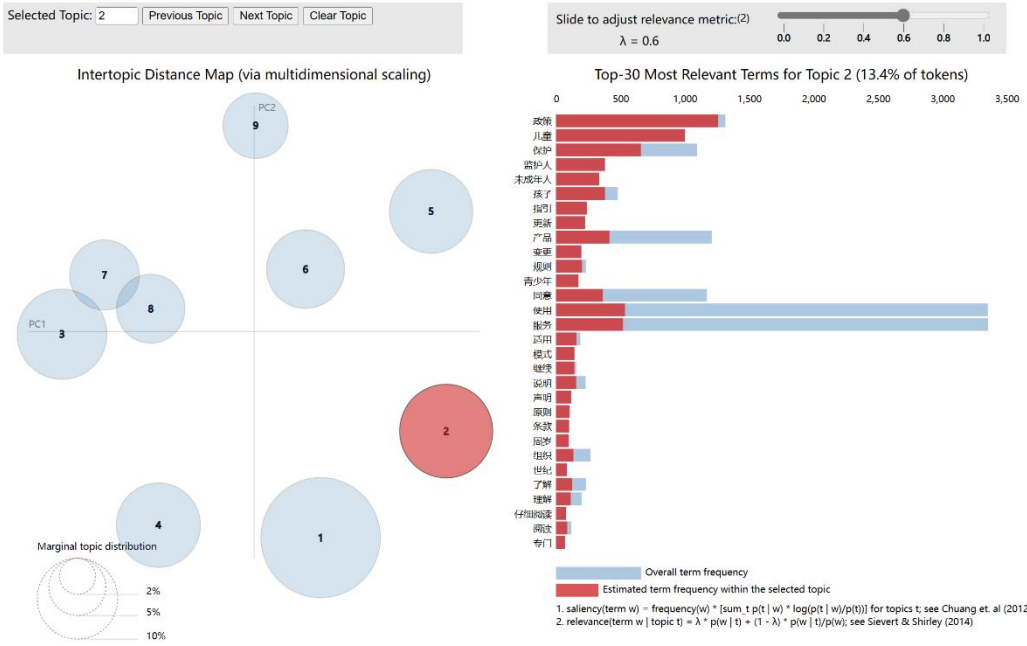
3.LDA主题识别

本研究采用困惑度（Perplexity）与一致性（Coherence）指标确定LDA主题数。

当主题数9时且LDA模型迭代次数为500时，困惑度数值最小、一致性数值最大同时各个主题间图像重叠区域最小，故确定LDA模型的迭代次数为500主题数为9。



序号	主题解释	前5个主题权重最高词组
1	信息获取与存储	服务，使用，提供，功能，收集
2	特殊信息管理	儿童，保护，使用，服务，产品
3	隐私政策概述	账号，认证，登录，帐号，身份
4	信息安全（Cookie技术）	推荐，发布，使用，个性化，Cookie
5	信息共享（用户披露）	同意，公开，法律法规，共享，授权
6	信息安全（信息保护）	安全，保护，数据，保障，使用
7	信息安全（安全技术）	设备，网络，应用，识别，日志
8	信息共享（跨境转移）	删除，存储，记录，搜索，保存
9	法律规定	注销，事件，账号，安全事件，处置



pyLDAvis可视化结果

本研究将LDA模型生成的9个主题进行进一步归纳合并，最终整理出6个类别：隐私政策概述、信息获取与存储、信息共享、信息安全、特殊信息管理与法律规定。

1.完整性评价

以《网络安全法》为主进行编码，形成6个一级指标和20个二级指标，构建社交应用隐私政策完整性评价体系。

一级指标	二级指标	指标解释
隐私政策概述	政策可读性	隐私文本的长度、内容是否易于接受
	政策历史版本及修订通知	是否提供历史版本查阅和新版本修订提示
信息获取与存储	基本信息	明确应用名称、开发方信息
	信息收集	隐私信息如何在用户同意下被收集
	用户信息用途	用户隐私信息用途
	信息存储	用户隐私信息存储地域时间及方式
	信息删除与修改	信息在应用内如何被删除与修改
信息共享	用户信息披露	共享转让披露个人信息的情形
	个人信息跨境转移	信息跨境转移处理方法
	用户选择权力	用户是否能拒绝数据给第三方共享与个性化服务
	Cookies及同类技术使用	应用如何利用用户使用过程中产生的Cookies信息
信息安全	隐私安全技术保障	用于保护隐私信息的技术手段
	第三方SDK接入库	列举第三方SDK库接入信息
	安全事件应急处置措施	是否包括对安全事件应急处置的说明，制定应急预案、进行应急处置、提交报告并告知受影响用户
特殊信息管理	未成年人信息管理	如何对未成年人隐私进行有效保护
	老年及残障人士信息管理	是否说明老年及残障人士的信息收集、处理和保护规则
	遗产信息管理	自然人死亡后其隐私数据如何管理
	应用法律责任	解释应用承担法律责任情况

（1）隐私政策概述

有8款隐私政策生效时间和更新时间不一致，不能为用户提供重新考虑是否同意的时间。70%的社交应用为未成年人隐私政策、第三方SDK库和Cookie使用说明提供了单独编写的政策文本，并在主政策文本中提供了超链接跳转，方便用户更轻松地了解这些政策条款。

（2）信息获取与存储

各应用都详细说明了隐私信息的收集过程，其中有19款主要通过用户提交或使用服务中记录的方式来获取用户信息，其余则主要通过Cookie等技术或第三方获取。只有腾讯软件的微信、QQ，以及淘宝、京东在隐私政策中明确了敏感信息的收集方式。在存储方面，只有28%的社交应用明确了隐私信息的存储期限和存储地点。

（3）信息共享

所有的政策文本都列举了单独的章节，对用户信息在共享、转让和公开披露时涉及的信息流动风险进行了表述。其中75%的应用明确指出，在委托处理、共享、转让和公开披露用户个人信息时，开发商或运营商必须事先征求授权同意。

（4）信息安全

所有隐私政策中都提供了Cookie技术和第三方SDK接入库的文档，其中67%和57%的社交应用单独提供了Cookie技术和第三方SDK接入库的详细清单，并在隐私政策中加入超链接以方便阅读。但仅有32%的社交应用提供了隐私安全应急事件的处置方式，并在文本中明确了用户在此过程中所承担的法律责任。

（5）特殊信息管理

目前法律和伦理框架下对逝者个人信息的处理还存在一定的限制，但有Keep，抖音，豆瓣等5款社交应用进行了跟进。各隐私政策对未成年人的信息保护相对成熟，大多数社交应用面向未成年人提供服务，年龄界定大致相同。但在老年人和残疾人群体的隐私信息管理方面，所有社交应用都没有提供专门的政策进行保护。

（6）法律规定

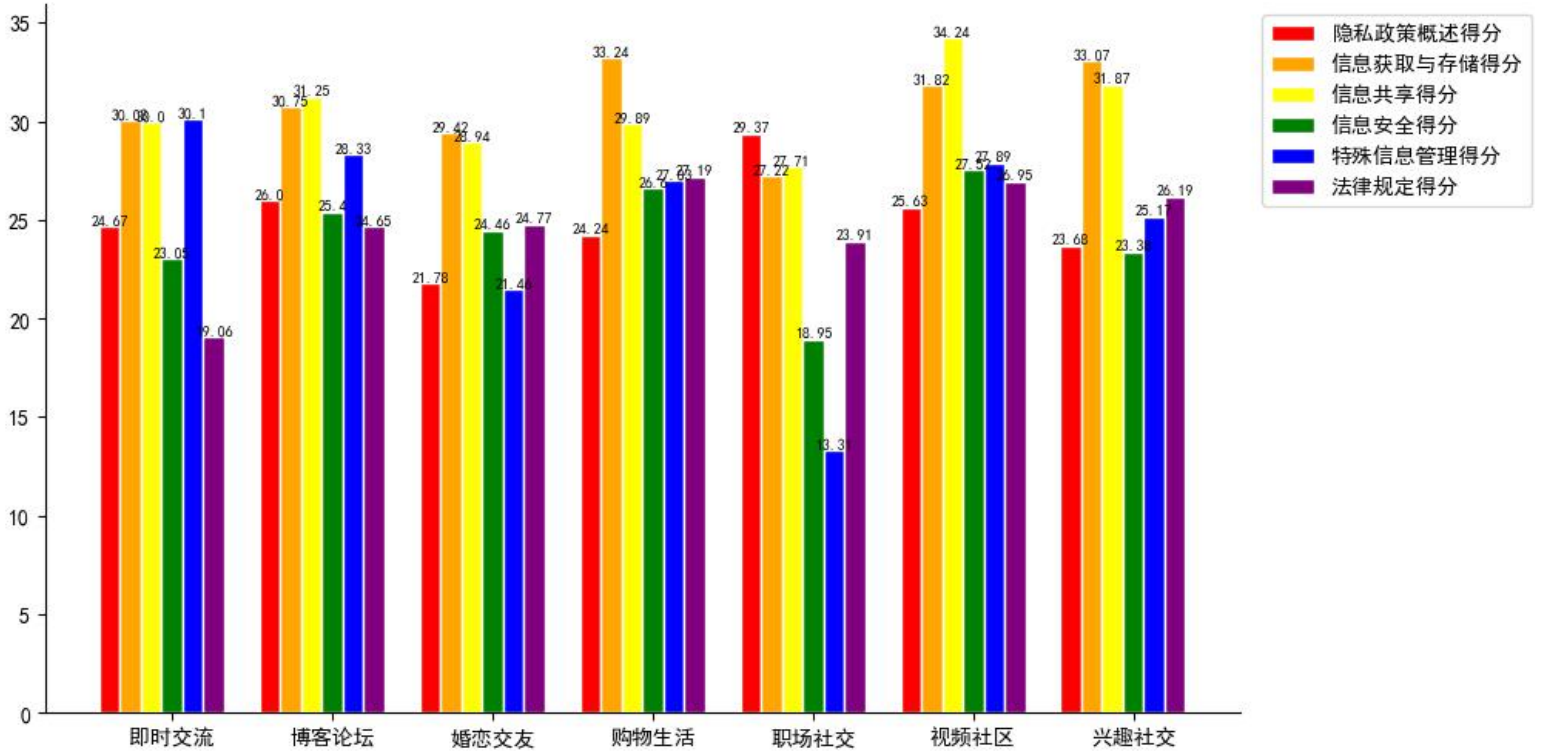
64%的社交应用针对法律框架下的用户权利进行了保障，用户有权要求社交应用提供收集的隐私信息清单，并提供信息修改和账号注销服务。所有的社交应用都在隐私政策最后提供了争议解决服务，但其中不到一半的应用明确了争议过程中用户和社交应用平台所需承担的法律责任，仅35%的社交应用提供了非常详细的争议处理方式，包括联系方式、联系地点、争议处理时间期限和验证通道等。

2.阅读感评价

首先利用文本可读性（readability）对政策文本进行可读性分析

$$readability = \frac{\alpha + \beta}{2}$$

其中 α 表示每个分句中的平均字数，β 为每个句子中副词和连词所占的比例。文本可读性指标越大，代表文本复杂度越高，可读性越差。

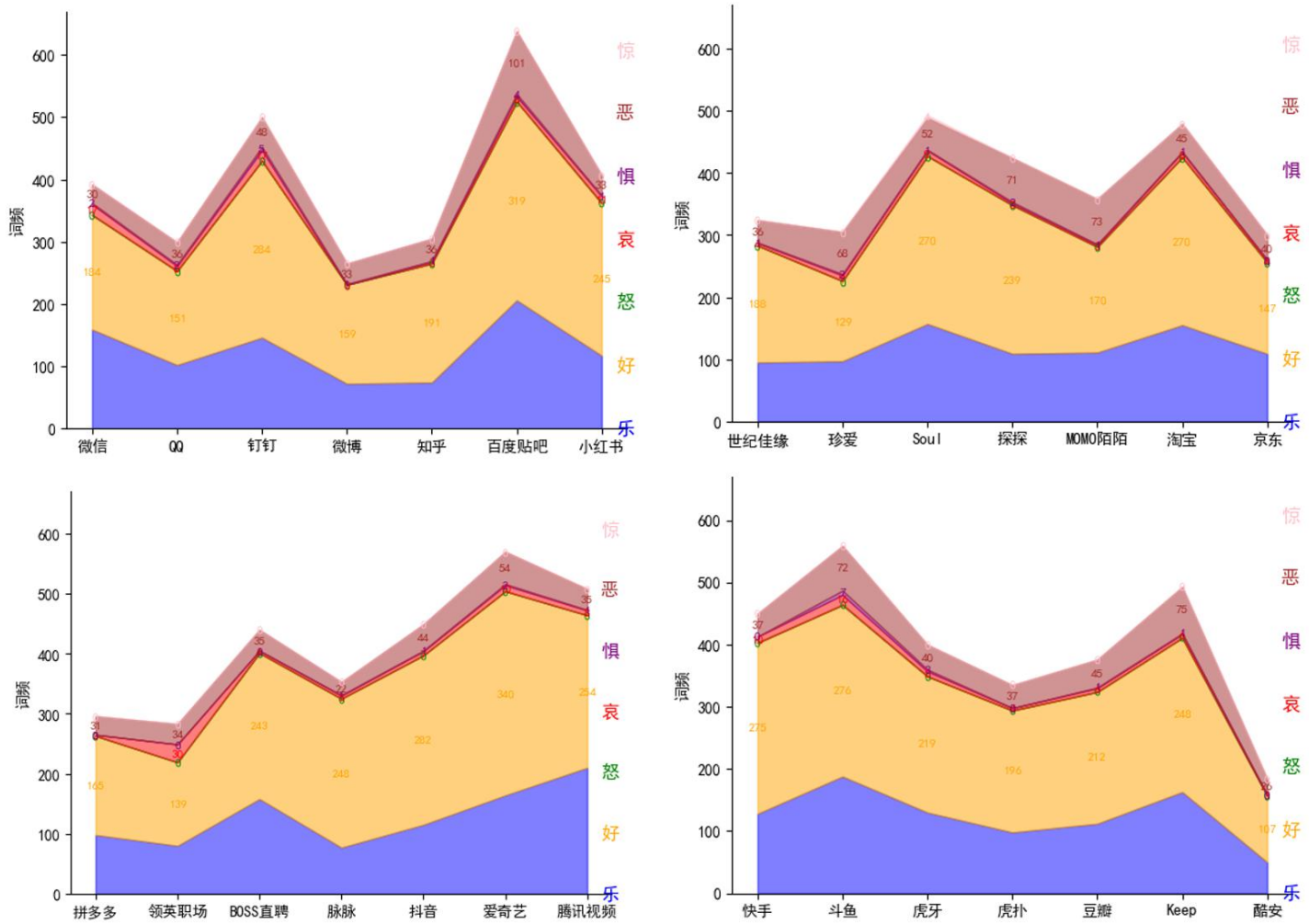


各类社交应用的文本可读性表现结果大致相同，其中信息获取与存储和信息共享章节的信息可读性普遍较差，特殊信息管理和法律规定得分章节的信息可读性普遍较好，这从侧面说明社交应用隐私政策在信息收集、存储与共享方面存在说明过于冗长，可解释性不高的问题，在特殊信息管理与用户权力和法律方面介绍过少，难以让用户快速了解隐私政策内容，培养用户的个人信息自我保护意识与权利意识。

再利用大连理工大学情感词典对相关指标的政策文本进行情感词语统计。由于隐私政策文本涉及大量专业术语，而现有情感分析领域多用于用户行为特征明显的语料库中，故本研究不针对得到的分类情感词语进行极性值计算，通过乐，好，怒，哀，惧，恶和惊7类情感词的分布情况对政策文本的情感倾向进行描述，

从图中可以看出社交应用隐私政策文本中类别为乐、好的词语占据了绝大部分，还有少量类别为惊的词语，除此以外，恶、惧、哀和怒类别下的词语几乎没有出现。

由此也能看出隐私政策文本在编写时尽量采用温和积极的语言，提高文字亲和力，但并未在文本中展现出爱憎分明的界限。



（1）加强法治建设和用户权利保障

社交应用应该在有关法律政策框架下建立个人信息获取与协商分级机制，充实用户权利内容，平衡应用平台与用户权利。

（2）加强对特殊群体的保护

对于老年人、残疾人以及遗产隐私信息，社交应用还需要明确具体的技术与管理双重保护手段。相关法律体系也需要不断推动改革，切实保障特殊全体隐私安全。

（3）加强政策文本可读性

政策文本中应合理控制篇幅，对各章节的内容进行简化，通过研发可视化的隐私政策，提升用户的理解能力加强对用户的隐私保护意识的引导，让用户了解隐私重要性，从而加强用户个人信息自我保护意识。

未来，将围绕社交媒体用户对于隐私政策的阅读意愿，用户的感知测度为研究重点，深入探讨不同类型的社交媒体APP的隐私政策合规性的用户阅读意愿，促进形成完善的、差异化的信息保护政策内容。

恳请各位专家和领导提出宝贵意见！

感谢聆听，敬请指正！