2024年海南省大学生

RPA+AI创新挑战赛

AI驱动的社交媒体热点追踪与分析系统 项目方案

学校

管他对不队参赛队编制

2024年11月12日

作品信息

参赛作品：AI驱动的社交媒体热点追踪与分析系统

**队伍名称**：管他对不队

**队伍成员**:

**队伍口号**：管他对不队，智胜未来！

**所属院校/专业**:

**作品应用场景**：实时追踪热点进行企业市场调研、用户需求洞察、内部沟通与协作、产品优化与创新……

**作品应用项目**：针对各类企业（尤其是消费品、科技、金融等行业）、市场调研公司、公关公司等，通过实时监测和分析社交媒体数据，帮助企业在市场调研、品牌管理、危机公关等方面做出更加精准的决策。其跨平台的特点和强大的数据处理能力，填补了现有市场的空白，具有广阔的商业前景。

**作品简介：**随着社交媒体的普及和数据的快速增长，RPA与AI技术在社交媒体分析中起到相辅相成的作用，扮演着越来越重要的角色。RPA技术能够自动化从不同社交媒体平台收集大量数据，而自然语言处理等AI技术，人工智能可以帮助分析社交媒体用户的行为模式，识别评论中的情感倾向，最后通过提供可视化图表，帮助企业和研究者做出更精准的决策。

**目 录**

[一、背景介绍 4](#_Toc182168522)

[二、需求分析 4](#_Toc182168523)

[三、建设方案 8](#_Toc182168524)

[四、方案的技术创新 17](#_Toc182168525)

[五、方案价值与收益 18](#_Toc182168526)

# 一、背景介绍

随着社交媒体的普及和数据的快速增长，社交媒体监测对于一个企业来说变得十分重要。社交媒体监测的必要性在于全面了解用户反馈、市场趋势及潜在危机。

而在当前的社交媒体环境下，企业面临着海量的数据和信息。传统的手动监测和分析方法已经无法满足快速响应市场变化的需求，企业需要一种能够自动化收集、分析和报告社交媒体热点的工具，以便及时了解市场动态、竞争对手的动向以及消费者的意见和情绪。

所以RPA技术能够自动化从不同社交媒体平台收集大量数据，包括用户评论、帖子、点赞等，为后续分析奠定基础。同时，人工智能技术在社交媒体分析中扮演着越来越重要的角色，情感分析可以识别评论中的情感倾向（积极、消极或中立），帮助企业掌握公众对品牌或产品的态度。

# 二、需求分析

## 概述

社交媒体作为一种新型的渠道，在客户企业所涉及的社交媒体环境下通常会出现以下痛点：

### 数据过载

社交媒体上每天产生大量的数据，企业难以从中筛选出有价值的信息。

### 实时性要求高

市场变化迅速，企业需要实时监控社交媒体动态，以便快速做出反应。

### 现有爬虫工具使用门槛高

许多现有的爬虫工具需要用户具备一定的编程知识和技术背景，这对于没有技术基础的营销人员和业务分析师来说是一个显著的挑战。他们往往需要依赖IT部门的支持，导致项目推进速度缓慢，且沟通成本高。

### 资源有限

专业的社交媒体分析师和数据科学家资源有限，且成本高昂。

### 部分信息获取工具抓取效率低下

传统的单线程爬虫工具在面对大规模数据抓取任务时，效率非常低下。它们无法充分利用现代计算机的多核处理能力，导致数据抓取过程耗时较长，无法及时获取最新的热点信息，影响企业对市场动态的快速响应。

而本项目可以有效地处理这些问题，并在以下场景中发挥有效的作用。

## 可应用场景概览

### 实时追踪热点进行企业市场调研

### （1）客户现有业务流程分析及客户需求

场景描述: 经典企业在推出新产品或服务时，需要了解市场趋势。通过本项目，企业可以实时监控各大社交媒体平台上的热点话题、用户讨论和情感倾向，从而制定更有效的市场策略。

### （2）项目特点

本项目通过实时抓取和分析数据，帮助企业快速响应市场变化。

### 用户需求洞察

### （1）客户现有业务流程分析及客户需求

场景描述: 通过分析社交媒体上的用户讨论和反馈，企业可以深入了解用户的需求和痛点，发现新的市场机会。这些数据可以为产品研发提供有价值的参考，帮助企业开发出更符合市场需求的创新产品。

### （2）项目特点

系统支持多平台（现已开发出微博，未来将会支持微信、抖音等）的数据抓取和分析，提供全面、统一的品牌声誉监控视图。

### 内部沟通与协作

### （1）客户现有业务流程分析及客户需求

场景描述: 企业内部不同部门可以利用该系统共享社交媒体热点数据，促进跨部门的沟通与协作。例如，市场部门可以与研发部门分享用户反馈，共同推动产品改进。

### （2）项目特点

处理过后的帖子内容文本数据和情感倾向数据在格式上是统一的，利于进行不同部门之间后续的数据再处理。

### 产品优化与创新

### （1）客户现有业务流程分析及客户需求

场景描述: 通过分析社交媒体上的用户评论和反馈，企业可以获取真实的用户体验数据，指导产品的优化和创新。例如，某款产品在特定功能上受到用户批评，企业可以根据这些反馈进行改进。

### （2）项目特点

这种基于用户反馈的产品迭代模式，有助于提升产品质量和用户满意度，增强市场竞争力。

# 三、建设方案

## 项目建设方案介绍

本项目用RPA技术多线程、自动化收集社交媒体数据，结合两种方法进行数据预处理，然后通过自然语言处理（NLP）技术分析文本内容，识别主题和情感倾向，最后将数据整合为多个饼状图展示给用户。

## 项目设计目标

### 自动化数据收集

目标: 利用RPA技术实现从多个社交媒体平台（如微博、微信、抖音等）自动化收集数据。

详细描述: 通过“影刀”平台搭建RPA机器人，通过半自动登录社交平台之后，抓取网络热点，以及用户发布的内容、评论等关键数据。

### 智能的文本预处理

目标：实现智能的文本预处理功能，以提高情感分析和数据挖掘的准确性和效率。

详细描述：筛选掉帖子内容中的#话题标签#；通过筛选字数较长且内容偏向于新闻报道的帖子，减少数据集中中立性内容的比例，使更多内容来源为用户生成的内容。

### 文本内容分析

目标: 使用自然语言处理（NLP）技术对抓取的文本数据进行深入分析。

详细描述: 应用NLP算法识别文本中的主题和情感倾向，提取关键词，并进行情感分析（如积极、消极、中立）。

### 数据可视化

目标: 提供直观的数据展示，帮助用户理解分析结果。

详细描述: 开发数据可视化工具，生成饼状图等图表，展示情感分布、主题变化等结果。

### 多线程支持

目标: 支持从多线程抓取和分析数据。

详细描述: 确保系统能够兼容多线程处理数据，不同子线程在进度上可以互相兼容。

### 用户友好的界面

目标: 提供易于使用的界面，方便非技术用户操作和查看分析结果。

详细描述: 设计简洁直观的用户界面，使用户能够轻松设置数据抓取任务、查看分析结果。

### 可扩展性与定制化服务

目标: 根据用户需求提供定制化功能和高级分析服务。

详细描述: 提供模块化设计，允许用户根据需要选择和添加高级功能模块，如深度分析报告、竞品对比等。

## 项目设计思路

本项目旨在利用RPA技术和AI算法自动化收集和分析社交媒体数据，以实现情感分析等结果。首先，通过RPA技术自动登录社交平台，多线程抓取用户发布的内容、评论等关键数据。接着，将抓取数据进行预处理并写入到数据表格中。然后使用自然语言处理（NLP）技术对处理后的数据进行深入分析，识别情感倾向，并进行情感分类。然后，开发数据可视化工具生成饼状图等图表，提供直观的数据展示，帮助用户理解分析结果。此外，整个系统设计注重用户友好性、多线程支持以及可扩展性，确保能够满足不同用户的定制化需求。

## 流程自动化风险与分析

### 主线程的流程自动化风险与分析

|  |  |  |  |
| --- | --- | --- | --- |
| **步骤** | **业务操作** | **风险分析** | **RPA操作** |
| 1 | 打开社交平台页面 | 风险点：系统响应不及时  解决方案：增加指令运行等待时间 | 打开社交平台页面 |
| 2 | 登录社交平台 | 风险点：（1）账号、密码错误。（2）网页是已登录状态，使登陆指令无效执行。  解决方案：（1）使用二维码登录方法，抓取二维码图像并添加基于条件语法的登陆确认页面，保证用户登录（2）通过基于网页元素检测的条件语句来跳过已经登陆的网站 | 登录社交平台 |
| 3 | 切换到热点新闻页面 | 风险点：鼠标模拟点击失效  解决方案：通过正则表达式检测对应的网页元素，使鼠标模拟点击更加精准 | 切换到热点新闻页面 |
| 4 | 抓取热点新闻 | 风险点：（1）抓取过多热点新闻，使程序运行时间过长（2）抓取的新闻无法参与后续的数据处理  解决方案：（1）减少热点新闻的抓取数量，将抓取范围限制为一页（2）使用数据表格以储存热点新闻的标题与链接 | 在热点新闻页面中循环相似元素为#热点标题#的对象，将热点新闻的标题与链接保存到数据表格的A列与B列中。 |
| 5 | 创建数个子线程，开始文本抓取与处理工作 | 风险点：（1）打开线程数过多，造成不必要的占用。（2）打开网页数过快，使IP被封禁。  解决方案：（1）设置打开的线程数为4。（2）每打开一个线程等待0.1秒。 | 并发调用子流程“数据抓取” |
| 6 | 等待子线程全部结束并进行下一步操作 | 风险点：（1）未等所有子线程全部抓取并处理完数据便继续主线程的进程。（2）最后一个线程抓取最后一个新闻标题相关帖子并分析情感倾向工作时，主线程会在情感倾向分析完毕并写入之前继续运行。（3）进行条件循环时会一直进行循环检测，占用过高。  解决方案：（1）设置条件语句和子线程对数据表格的写入，从而进行检测是否全部结束。（2）在检测到最后一行“完成读取！”后等待0.2秒。（3）每次进行循环时等待0.1秒 | While条件循环：如果数据表格最后一行B列为“完成读取！”，则继续主线程的操作，否则等待0.1秒重复检测。 |
| 7 | 数据读取与翻译 | 风险点：逐行读取表格内容，会将文本内容计入情感倾向数据中  解决方案：通过设置一个递增值，只读取偶数行的情感倾向数据，提升运行速度。 | 逐行读取数据表格内容，将所有热点新闻的标题储存到列表，并开始遍历，将每个热点新闻的帖子对应的情感倾向以及出现次数写入到字典，最后将字典内容汇总到列表中。 |
| 8 | 生成饼状图 | 风险点：（1）N张饼状图分成N次输出，不符合界面友好的特点。（2）输出的图标不支持中文标题和变量值。（3）排版、饼状图间隔、滚动条、区域颜色等页面设计问题。（4）若读取的情感倾向出现次数为0，仍会在图表中渲染。  解决方案：（1）通过遍历将N个子图渲染到一个页面中（2）通过Matplotlib内置函数预设字体解决。（3）通过调整Matplotlib设置图标参数、在页面渲染中设置分栏、将页面嵌入到Tkinter窗口中、提前预设情感倾向对应的区域十六进制颜色代码解决。（4）在读取数据中过滤掉值为0的键值对 | 通过使Python程序读取上一步的两个列表变量，来生成图表 |

### 子线程的流程自动化风险与分析

|  |  |  |  |
| --- | --- | --- | --- |
| **步骤** | **业务操作** | **风险分析** | **RPA操作** |
| 1 | 获取数据表格总行数 | 风险点：无  解决方案：无 | 读取数据表格总行数 |
| 2 | 循环读取数据表格内容（热点新闻标题+网址） | 风险点：无  解决方案：无 | 通过行数循环，读取表格中热点新闻的标题与网址 |
| 3 | 打开新闻标题对应的网址 | 风险点：无  解决方案：无 | 打开网址 |
| 4 | 逐项抓取热点新闻的相关帖子与评论内容 | 风险点：（1）无法定位到新建的标签页（2）抓取的帖子数量过多  解决方案：（1）通过保存网页对象使后续操作精准定位（2）在牺牲少量数据精度的前提下限制每个热点新闻抓取的帖子数量为5 | 在热点新闻列表中遍历每个热点新闻的数据，打开其标题对应的链接，并对其相关帖子内容的文本进行抓取，并保存到临时变量之中。 |
| 5 | 数据预处理 | 风险点：（1）每个帖子内容中含有话题标签，影响后续对其情感的分析精确度（2）帖子内容有部分是完全中立的新闻类内容，相对之下不具有参考价值。  解决方案：（1）通过删除抓取文本字符串中的热点新闻标题字符串，使内容情绪取向更加精准。（2）以文本长度为过滤参数，过滤掉大于100字的文本内容，降低新闻类内容的含量，同时利于情绪分析 | 将帖子的文本内容进行预处理，删除字符串中的子串（标题的标签字符串），并获取文本长度，以此进行过滤筛选。 |
| 6 | 文本情感倾向分析 | 风险点：无法获取要分析的文本内容  解决方案：精准设定不同变量，将用于文本情感分析的变量值对应到数据预处理后的帖子文本中 | RPA机器人调用NLP服务进行情感分析 |
| 7 | 数据写入 | 风险点：新写入的数据会覆盖原有数据表格的前两列（热点新闻标题、对应链接）  解决方案：从第三列开始写入。 | 将抓取的预处理后的数据与分析出的情感倾向从数据表格的第三列开始逐列写入 |
| 8 | 关闭网页 | 风险点：只关闭一个开启的网页。  解决方案：将关闭网页指令定位到每抓取完一条热点新闻之后。 | 每抓取完一条热点新闻并完成6-8步后关闭对应网页 |

## 机器人流程设计图展示与分析

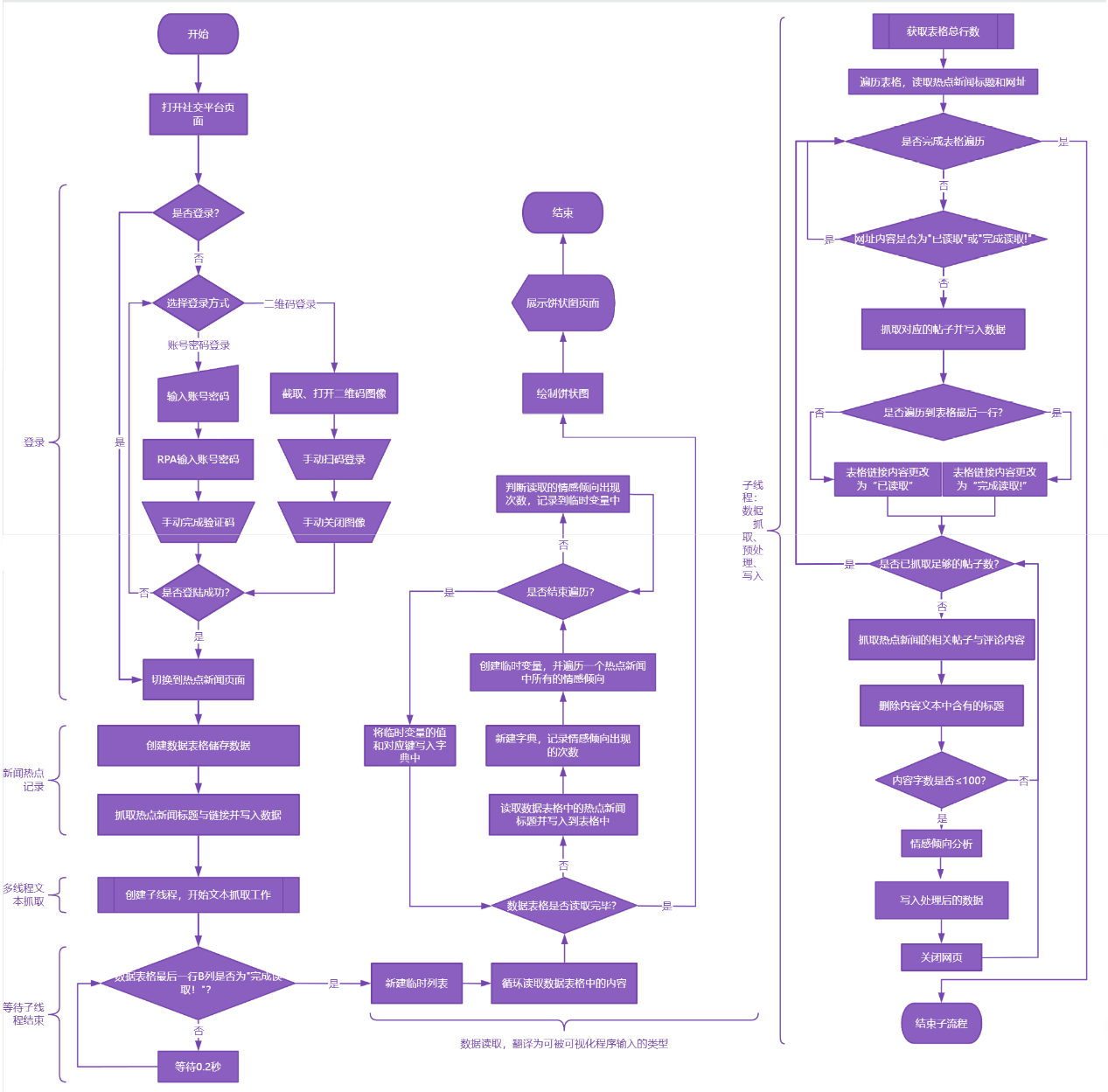


图1 业务流程图

## 流程步骤说明

### 主流程的步骤说明

|  |  |  |
| --- | --- | --- |
| **步骤** | **流程描述** | **机器人或人工** |
| 1 | 打开网址：https://weibo.com/hot/mine | 机器人 |
| 2 | 检测是否已经登录，若已经登录，则跳过3-7步，直接从第8步开始执行 | 机器人 |
| 3 | 点击登录，弹出登录信息输入界面，打开选择对话框并引导用户选择登陆方式 | 机器人 |
| 4 | 用户选择登陆方式 | 人工 |
| 5-1.1 | 如果用户选择的是二维码，则创建临时文件夹，截取并保存二维码图像，打开二维码图像，并提示用户扫码登录微博 | 机器人 |
| 5-1.2 | 用户扫码登录微博 | 人工 |
| 5-2.1 | 如果用户选择的是账号密码登录，则切换到账号密码登录选项，提示用户输入账号密码 | 机器人 |
| 5-2.2 | 用户输入账号密码，并完成人机验证。 | 人工 |
| 6 | 弹出选项，确认用户是否登录成功，若成功则继续，不成功则从第4步开始重新执行 | 机器人 |
| 7 | 用户选择是否登陆成功 | 人工 |
| 8 | 保存登陆成功后的网页对象，并切换到热点新闻-要闻界面，并设置整数变量count | 机器人 |
| 9 | 遍历并读取要闻界面中的热点标题，将标题内容与链接分别写入到数据表格的(count,A)与(count,B)中,每写入一次将count的值加1 | 机器人 |
| 10 | 通过For循环，间隔0.1秒并发调用1次子流程——“数据抓取”，共调用4次。 | 机器人 |
| 11 | 通过While循环，检测数据列表最后一行的B列数据是否为“完成读取！”以检测子流程是否全部运行完毕，若是，则退出循环进行下一步操作；若否，则等待0.2秒后重新检测。 | 机器人 |
| 12 | 新建列表title\_list，储存所有的热点新闻标题；新建列表dict\_emo\_list，储存所有键为情绪倾向名字；值每行数据表格的键出现的次数的字典 | 机器人 |
| 13 | 对数据表格进行遍历，按行读取数据表格的内容，将A列的数据写入列表title\_list中 | 机器人 |
| 14 | 创建三个变量，分别用于计数三种情绪倾向在一个新闻标题（每行数据表格）中出现的次数。并新建键为情绪倾向名字；值每行数据表格的键出现的次数的字典dict\_emo | 机器人 |
| 15 | 从第4列开始遍历，递增值为2，统计只会在偶数列出现的情绪倾向出现的次数 | 机器人 |
| 16 | 设置字典dict\_emo的键值对，并插入到列表dict\_emo\_list中 | 机器人 |
| 17 | 将变量dict\_emo\_list与title\_list作为饼状图的数据来源和表头输入到饼状图绘制程序中 | 机器人 |
| 18 | 渲染饼状图页面并对用户进行展示 | 机器人 |
| 19 | 用户关闭页面，结束 | 人工 |

### 子流程的步骤说明

|  |  |  |
| --- | --- | --- |
| **步骤** | **流程描述** | **机器人或人工** |
| 1 | 按行遍历数据表格，并读取B列内容 | 机器人 |
| 2 | 检测B列内容是否为“已读取”或“完成读取！”，如果是，则按下一行重新开始第一步；如果否，则将B列内容修改为“已读取”，并继续子流程 | 机器人 |
| 3 | 打开B列内容的网址，并保存为网页对象HS\_Page | 机器人 |
| 4 | 切换到“实时”帖子分类界面，设置变量count0=3,为后续写入数据表格的列数 | 机器人 |
| 5 | 遍历抓取帖子内容文本 | 机器人 |
| 6 | 进行第一步数据预处理：删除掉抓取文本的字符串含有的话题文本（#新闻标题#），并返回结果 | 机器人 |
| 7 | 进行第二步数据预处理：获取第一步预处理后的文本长度，并检测长度值是否小于等于100.如果否，则返回第六步重新抓取；如果是，则进行下一步操作 | 机器人 |
| 8 | 将最终的文本写入遍历的行数，列数为count0，并将count0加1 | 机器人 |
| 9 | 通过NLP进行情感倾向分析，将返回的结果写入遍历的行数，列数为count0，并将count0加1 | 人工 |
| 10 | 检测count0的值是否大于等于13，如果是，则退出抓取帖子内容文本的遍历，同时关闭网页；如果否，则返回第六步继续抓取。 | 机器人 |
| 11 | 检测遍历的行数是否为最后一行，如果是，则将B列内容修改为“完成读取！”，并结束子流程 | 机器人 |

# 四、方案的技术创新

## **自动化数据收集**

系统应能自动从各大社交媒体平台收集数据，减少人工干预。

## **实时数据分析**

提供实时的数据分析功能，帮助企业快速响应市场变化。

## 智能数据预处理

通过数据整理和过滤实现智能的文本预处理功能，以提高情感分析和数据挖掘的准确性和效率。

## 深度洞察生成

利用AI技术进行情感分析、趋势预测等，提供深度的市场洞察。

## 用户友好的界面

简洁直观的用户界面，便于非技术人员也能轻松使用。

## 定制化报告

根据企业的具体需求，生成定制化的报告和仪表板。

## 多线程集成

支持多线程的数据采集和分析，倍数提升运行效率。

# 五、方案价值与收益

## （一）方案给客户带来的收益和价值

### 提高效率

人工操作时间 vs RPA操作时间：通过自动化技术，可以显著减少手动操作的时间。例如，原本需要数小时的数据处理工作，RPA机器人可以在几分钟内完成，从而大幅提高工作效率。

处理数据量：RPA能够处理大量数据，远超人工能力范围，使得数据分析更加全面和深入。

### 提升准确性

错误率降低：人工操作容易出现疲劳和失误，而RPA机器人可以24/7不间断工作，且几乎不会出错，大大提升了数据处理的准确性。

数据一致性：自动化流程确保每次操作都按照相同的标准进行，避免了人为因素导致的数据不一致问题。

### 降低成本

人力成本：减少了对人力资源的依赖，降低了长期的人力成本。

运营成本：自动化流程减少了手动操作所需的时间和资源，从而降低了整体运营成本。

### 改善用户体验

客户满意度：通过提供更准确和及时的服务，提升了客户满意度。

品牌形象：高效的业务流程和优质的服务有助于提升品牌形象。

## （二）、方案的可推广性说明

### 跨行业适用性

广泛适用：该方案适用于各类企业的社交媒体数据分析，具有广泛性。

模块化设计：方案采用模块化设计，可以根据不同行业的需求进行调整和定制。

### 易于集成

兼容性强：RPA技术具有良好的兼容性，可以轻松集成到现有的IT系统中，无需大规模改造现有架构。

标准化接口：抓取的数据提供标准化的变量形式，方便与其他系统对接。

### 扩展性强

灵活扩展：根据业务需求的变化，可以灵活增加或减少RPA机器人的数量，以适应不同的工作量。

### 成功案例

测试反馈良好，进一步验证了方案的可行性和实用性。

### 总结

通过以上分析可以看出，该方案不仅能够为客户带来显著的收益和价值，还具备良好的可推广性和适应性，能够在不同行业中广泛应用。