--数据科学导论大作业--

# 解构对话：一份微信聊天记录的深度剖析

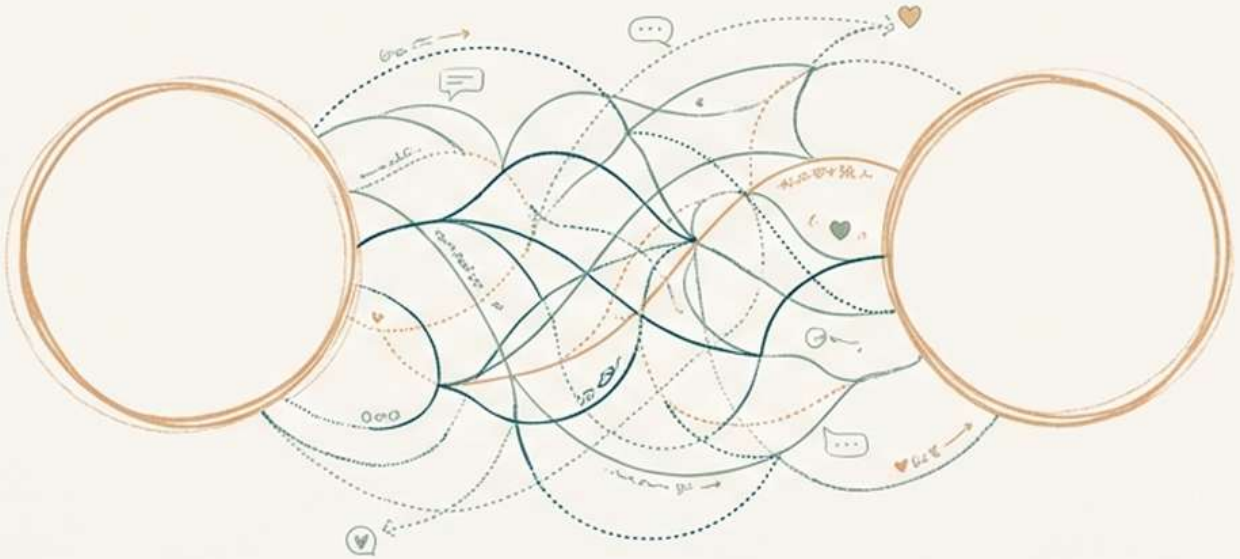
结合 Python 数据科学与 AI 洞察，揭示近万条消息背后的关系图谱

吴晨曦 10242150443

# 缘起：一份定制的“关系年度报告”

通用平台年度回顾

**每年，各大平台都会为我们生成年度报告。我们能否为自己最最重要的一段对话，也生成一份独特的“关系报告”？**



独特的二人“关系报告”

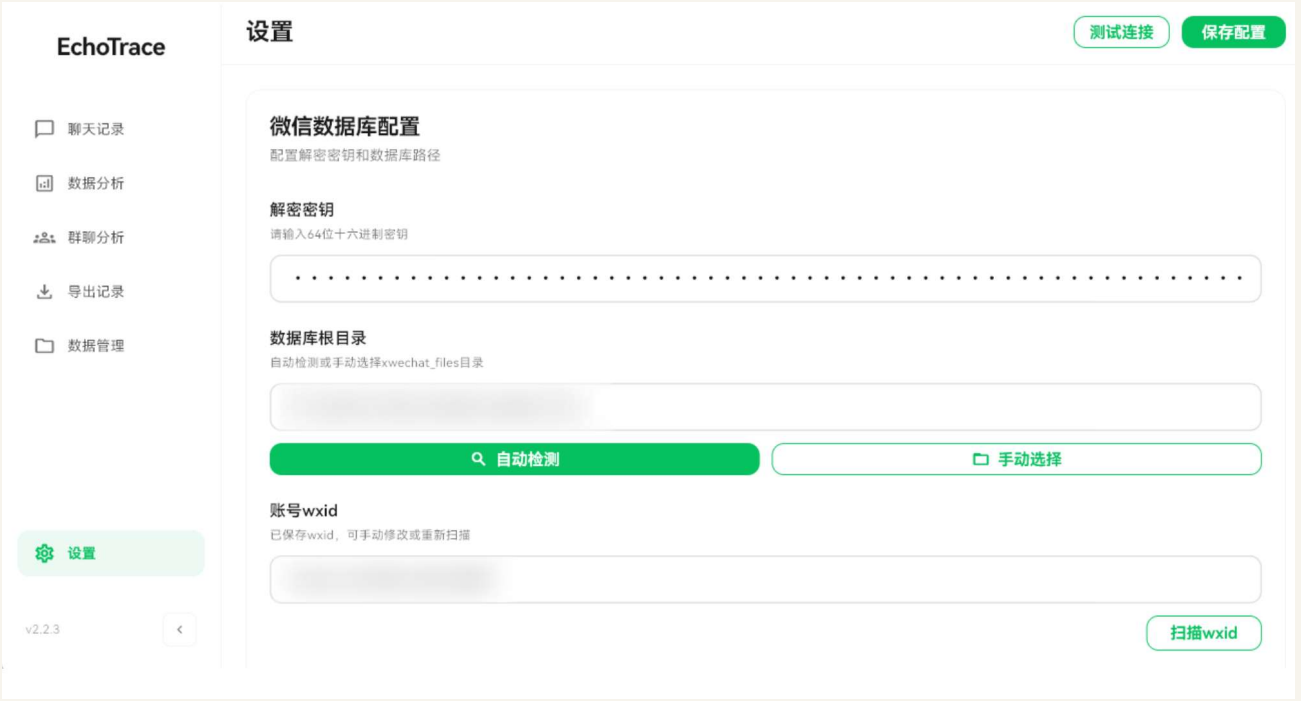
# 分析蓝图：从数据到洞察的六步法

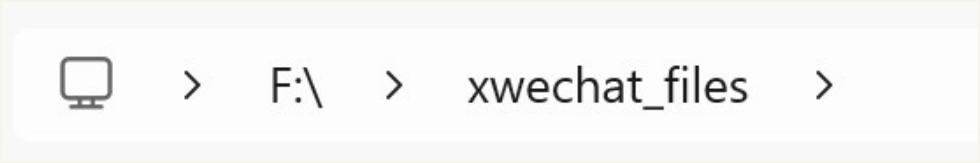
1. **数据准备**
   从海量记录中提取结构化信息。
2. **基本画像**
   勾勒对话的宏观轮廓。
3. **活跃节律**
   探寻沟通的时间模式。
4. **核心话题**
   解读我们最关心的事。
5. **情感动力学**
   量化感受的起伏与温度。
6. **AI 综合研判**
   生成最终的定性洞察报告。

# 0.数据导出——EchoTrace开源工具

- 原理：电脑版微信的聊天记录存储在本地的 SQLite 数据库中(加密)。密钥通常存储在微信运行时的内存中。
- 方法：通过开源工具（如 EchoTrace）解密、配置并以指定格式导出

[echotrace/docs/beginner_guide.md at main · yccccccy/echotrace]

# 1.数据预处理

## 一切分析始于干净、可靠的数据。

**数据源**

使用 EchoTrace 工具安全导出本地微信聊天记录为 JSON 格式。

**技术栈**

核心使用 Python 及 Pandas（数据处理），Matplotlib/Seaborn（可视化），Jieba（分词），Transformers（情感分析）等库。

**关键处理**

将 9,961 条原始消息转化为结构化的 DataFrame，提取时间、发送者、消息类型等关键特征，并进行数据清洗。**（转换时间戳、提取特征、删除空数据等）**


```
{
  "localId": 1646,
  "createTime": 1748786290,
  "formattedTime": "2025-06-01 21:58:10",
  "type": "文本消息",
  "localType": 1,
  "content": "哈哈哈哈哈哈啊哈哈哈哈哈",
  "isSend": 0,
  "senderUsername": "wxid_2gbxgakagb9422",
  "senderDisplayName": "myfriend",
  "source": "<msgsource>\n    <sec_msg_node>\n
},
{
  "localId": 1645,
  "createTime": 1748786305,
  "formattedTime": "2025-06-01 21:58:25",
  "type": "文本消息",
  "localType": 1,
  "content": "试试别的软件",
  "isSend": 0,
  "senderUsername": "wxid_2gbxgakagb9422",
  "senderDisplayName": "myfriend",
  "source": "<msgsource>\n    <sec_msg_node>\n
},
```

# 2.基本画像

## 对话画像：谁在主导这场对话？

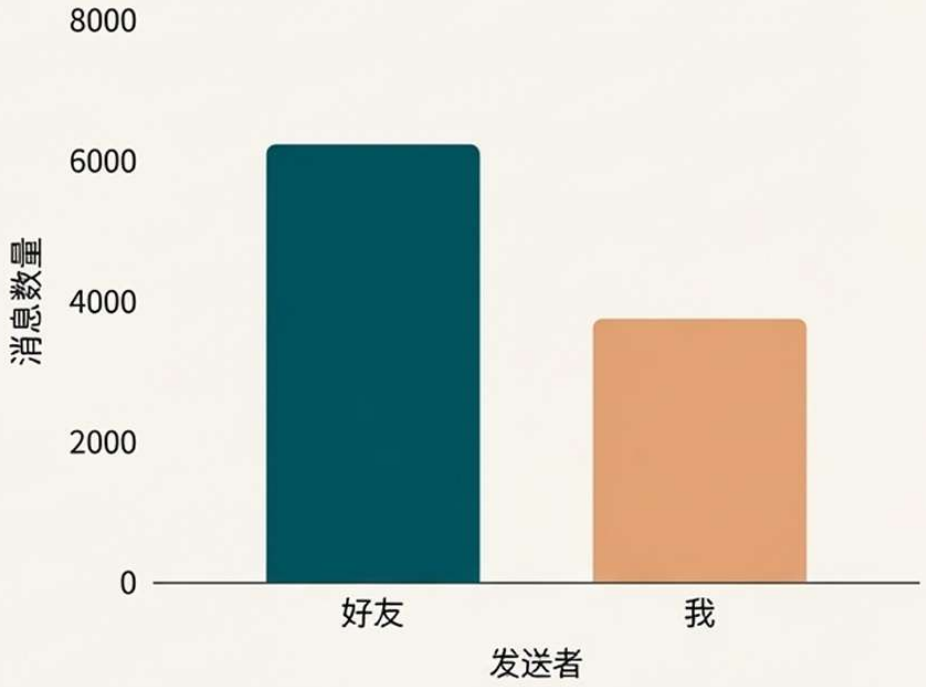
```
sender_counts = chat_df['sender'].value_counts()
```

**数据呈现清晰的“分享-回应”动态，而非均等的对话量。**

总消息数：**9,961 条**

**好友：6,204 条（62.3%）** - 主要的分享者与话题发起者。

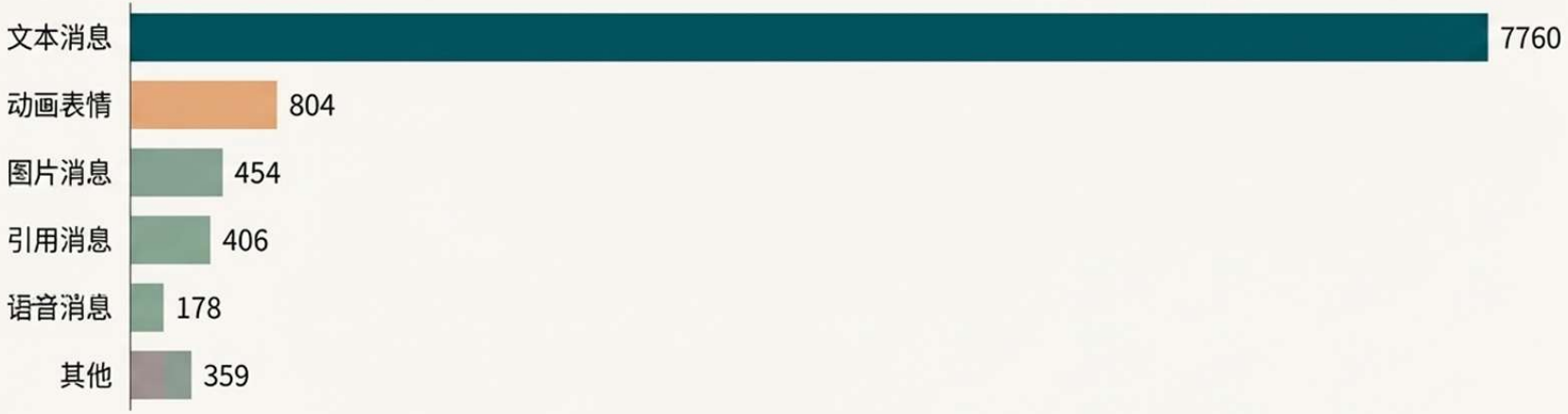
**我：3,757 条（37.7%）** - 积极的倾听者与回应者。



# 2.基本画像

## 沟通方式：不止于文字

```
message_type_counts = chat_df['message_type'].value_counts()
```




**文本消息**（78%）：沟通的主体，是深度分析的基础。

**动画表情**（8%）：情感的快捷方式，传递即时感受。

**图片消息**（4.6%）：分享生活瞬间，连接彼此世界。

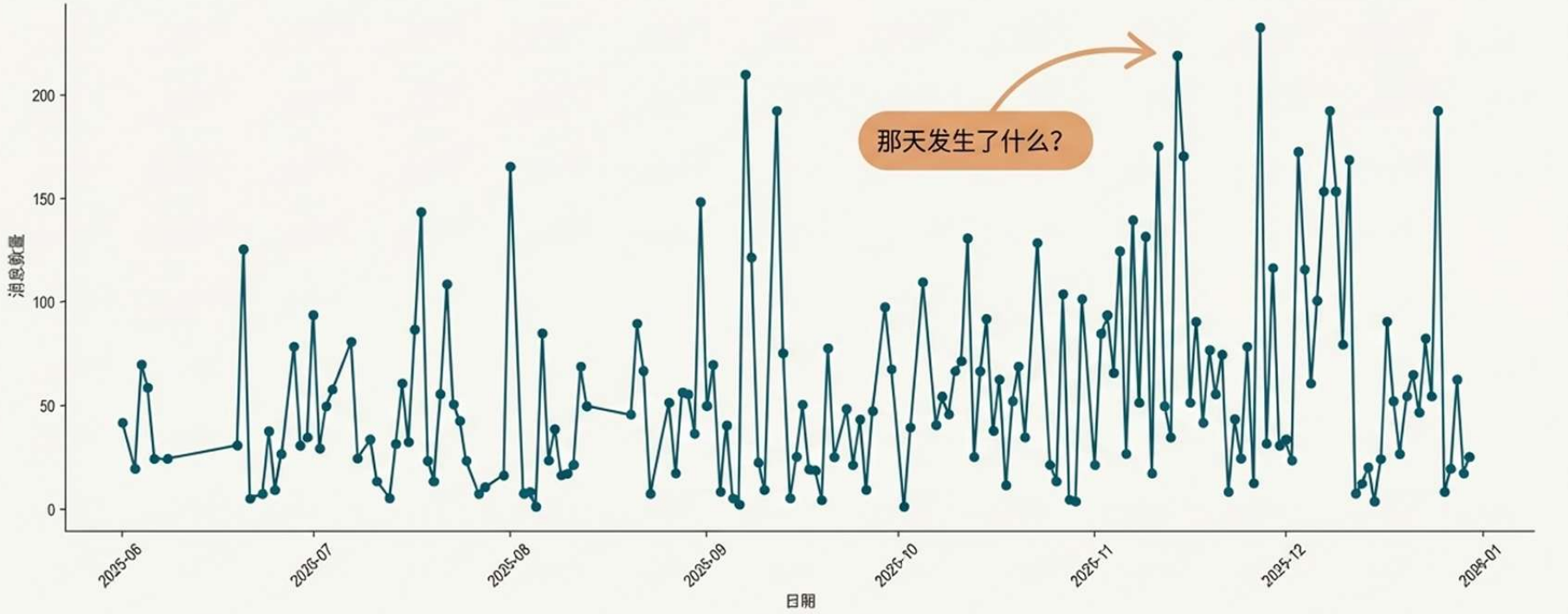
**其他：**引用、语音、链接等消息类型进一步丰富了交流的维度。

# 3.活跃节律

## 沟通的潮汐：每日消息量趋势

```
daily_trend = chat_df.groupby('date').size()
```

沟通强度并非一成不变，而是呈现显著的“峰谷”波动。这与双方的现实生活事件（如假期、工作繁忙期、共同经历）高度相关。

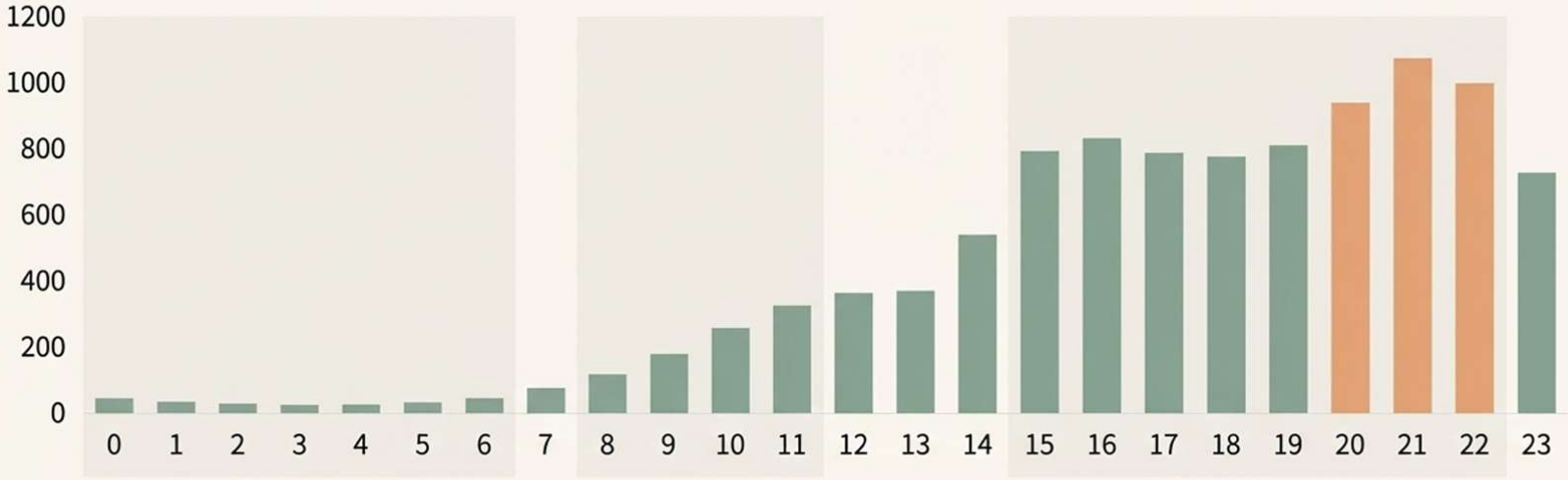


# 3.活跃节律

## 我们的“聊天时区”：每小时活跃度解密

```
hourly_activity = chat_df.groupby('hour').size()
```

对话拥有非常清晰的日内节律，集中在下午和夜晚，晚间 9-10 点是沟通的黄金时段。




**凌晨（0-7点）**：静默区，符合普遍作息规律。

**上午（8-12点）**：缓慢启动。

**下午至夜晚（15-23点）**：核心活跃区，反映了学生或上班族在日间的空闲时段及下班/下课后的放松时段进行集中沟通的习惯。

# 4.核心话题

- stop_word机制
- 结巴分词
- 词云展示

高频词汇揭示了对话的核心：紧密围绕“日常生活”、“个人感受”和“共同经历”展开。词汇的“去事件化”与“情感化”特征显著，表明对话价值在于持续的情感确认与联结。


**生活日常**

哈哈、吃、同学、老师、学校、回家

**情感表达**

感觉、真的、喜欢、开心、没事

**时间与状态**

今天、现在、晚上、回来

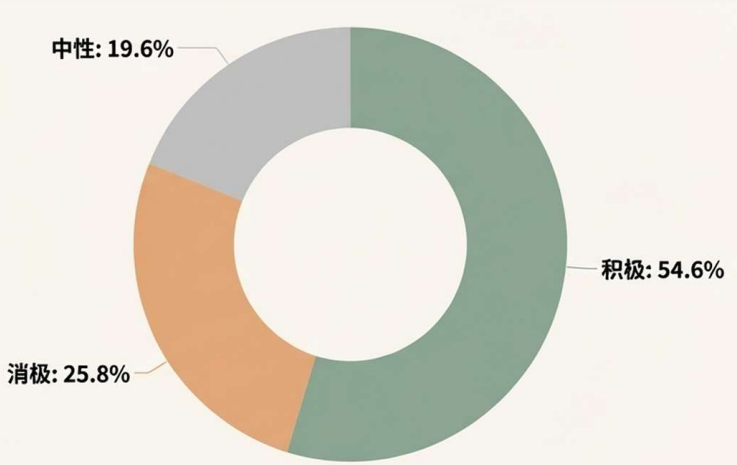
# 5.情感分析

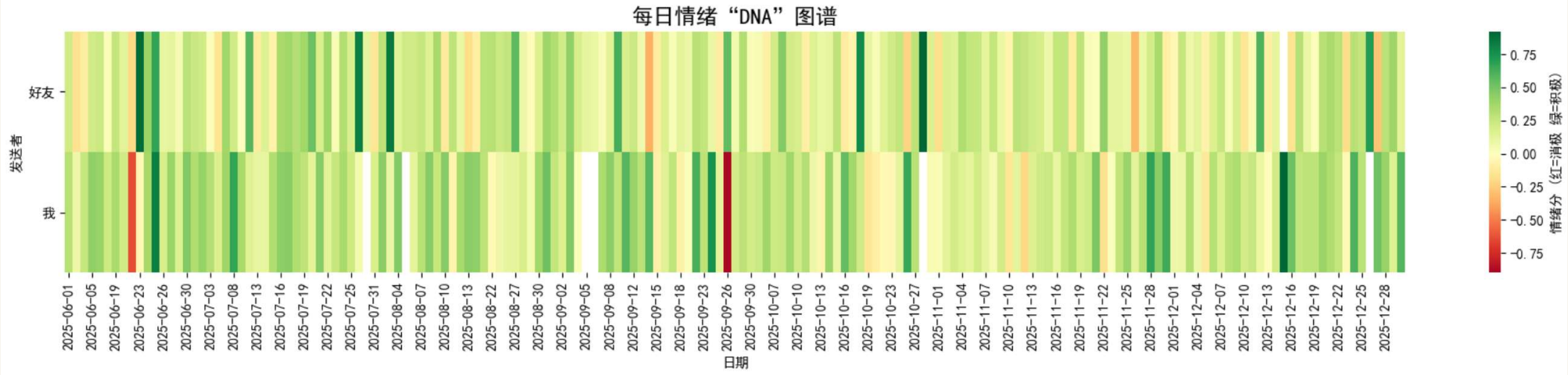
- 基于 **Transformer 架构**、利用**电商语料微调**的**二分类**情感分析系统

- **RoBERTa** (Robustly optimized BERT approach)
  - BERT 的进阶版
  - Transformer 架构：利用**自注意力机制 (Self-Attention)** 捕捉文本中字与字之间的上下文关系
- **后处理**
  - 中性区间过滤：0.4-0.6→归为中性
  - 概率线性映射：[0,1]→[-1,1]

# 5.情感分析

- 情绪分布饼图
- 每日情绪DNA
- 情绪波动曲线



# 情绪波动曲线：追踪高光与低谷时刻

通过量化每日情绪，我们可以精准定位关系中的“开心巅峰”与“情绪低谷”，并回溯其背后事件。

## Case Study

**日期** 2025-06-21，被标记为气氛[illegible]压抑的一天（平均分: -0.35）。

**触发消息** 真的笑鼠我了

**发送者** [illegible]

**分析** 这句话在[illegible]境中表达的是极度的幽默和开心，但模型因未能理解网络俚语“笑鼠”的正面含义，将其误判为强烈的负面情绪。**这个“错误”恰恰是理解模型局限性的一个绝佳案例。**

# 6.AI分析报告

- 准备文本内容进行分词
- 构建prompt
- 调用API（deepseek）

```
    prompt = f"""
    请根据以下聊天数据摘要，生成一份简短、深刻的中文分析报告。报告应包括：
    1. 对整体聊天风格和氛围的总结。
    2. 对主要参与者的沟通模式的简要分析。
    3. 基于高频词汇，推断出他们可能的核心话题或关系特点。
    4. 提出一两个有趣的观察或结论。

    **数据摘要:**
    - **时间跨度**: {duration_days} 天
    - **总消息数**: {total_messages} 条
    - **参与者消息分布**:
"""
```

**聊天数据分析报告**

**1. 整体聊天风格与氛围**

在长达212天的持续互动中，两位参与者共交换了近万条消息，形成了**高频、紧密且高度日常化**的沟通模式。聊天氛围整体呈现出**轻松、信任、生活化**的特点，双方建立了稳定的分享习惯。好友在消息数量上占据主导（62.3%），显示出更强的表达欲和分享主动性，而"我"则扮演了积极的倾听与回应者角色。这种不对等的消息量并未破坏平衡，反而可能暗示了一种**默契的互补关系**。

**2. 主要参与者沟通模式分析**

- **好友（主导分享者）**：其沟通模式具有**高频率、细节丰富、情感外露**的特点。占据六成以上的消息量，表明其习惯于将日常琐事、即时情绪与想法进行持续性输出。这种模式可能源于强烈的表达需求、对关系的信任，或希望通过分享来维系亲密感。
- **"我"（积极回应者）**：虽然消息量较少，但近38%的占比在长期对话中仍属高度参与。其模式更偏向**倾听、承接话题、提供情感反馈或实用建议**。这种"回应型"沟通维系了对话的流动性，并可能为好友的分享提供了稳定的情感容器。

**3. 高频词汇揭示的核心话题与关系特点**

从高频词分析，对话核心紧密围绕**两人共同的现实生活与情感世界**：

- **生活实景（"今天"、"吃"、"哈哈"、"回家"、"睡觉"）**：词汇极具日常性与即时性，表明聊天是彼此**生活节奏同步与陪伴**的重要方式。分享三餐、作息、日常趣事（"哈哈"）构成了关系的基底。
- **关系指向（"你"、"我"、"宝宝"）**：人称代词的高频使用凸显对话的**直接性与互动性**。"宝宝"一词的出现（即便频率不高）是重要信号，可能暗示着**亲密昵称、恋人关系或深厚的情感依恋**，这是定义关系性质的关键词。
- **情感与状态（"感觉"、"现在"）**：注重分享当下心境与感受，沟通具有**情感流动与状态同步**的特点，而非单纯事务性交流。

**核心推断**：两人很可能处于**亲密关系（如恋人或挚友）**中，**聊天主要功能是**维系日常情感连接、分享生活实感、提供陪伴**。关系已深度融入彼此日常生活。

**4. 有趣的观察与结论**

1. **"分享即陪伴"的非对称平衡**：尽管消息量有显著差距（62.3% vs 37.7%），但这并未导致关系失衡，反而可能形成了一种**稳定的"倾诉-承接"动态**。好友通过高频分享获得陪伴感与情感释放，"我"则通过积极回应获得参与感与关系满足。这挑战了"消息量均等才健康"的简单假设，体现了关系模式的多样性。
2. **词汇的"去事件化"与"情感化"**：Top词汇中罕见具体事件、人物或宏大话题，几乎全部聚焦于**个人化的瞬间感受与微观生活**。这表明，对话的核心价值不在于信息交换，而在于**持续的情感确认与联结**。他们的关系可能更依赖于这种日常"微循环"来滋养，而非依赖重大事件或深度讨论。

---

**总结**：这是一段深度嵌入日常、以情感陪伴为核心功能的亲密对话。其生命力不在于话题的波澜壮阔，而在于日复一日的分享与回应所构筑的稳定感和亲密感。

# 结论：一份数据驱动的“关系说明书”

本项目成功将海量、非结构化的聊天数据，转化为一幅关于沟通模式、互动节律与情感动态的结构化、可解读的洞察图谱。

---

数据分析不仅是技术工具，更提供了一种全新的视角，帮助我们回顾与理解自己的人际互动，珍视那些在日常交流中悄然流逝的情感与时光。

# 谢谢观看！