- 1. 开场白：一句话点明主旨
- 2. 第一部分：左侧——数据准备（打地基）
- 3. 第二部分：中间核心——双重算法并行（双保险）
- 4. 第三部分：集成与迭代（滚雪球）
- 5. 总结升华

这张图展示的是一个**“基于种子词扩展的领域情感词典构建流程”**。它的核心逻辑是：**利用少量的已知情感词（种子词），在大量文本数据中“滚雪球”式地挖掘出更多未知的情感词。**

为了让你在答辩或展示时讲得逻辑清晰且有深度，建议按照**“输入处理 → 核心算法（双轨并行） → 迭代优化 → 最终输出”**的逻辑线来讲解。

以下是具体的讲解策略和话术建议：

1. 开场白：一句话点明主旨

话术：

> “大家请看这张图，这是本研究提出的基于社交媒体文本的情感词典构建框架。针对通用词典在特定领域（如社交媒体）覆盖率不足的问题，我们设计了一套**‘双重校验 + 迭代扩展’**的自动化构建方法。”

---

2. 第一部分：左侧——数据准备（打地基）

讲解重点： 强调数据来源的真实性和预处理的重要性。

- **图示对应：** 左侧的长条框（数据爬取 → 预处理 → 词性标注）。
- **逻辑解释：**
  - **来源：** 用户生成内容（UGC），即社交媒体数据，特点是口语化、新词多^1^。
  - **关键步骤：** 经过清洗后，我们重点进行了 **词性标注** 。
  - **Why？** 因为情感往往蕴含在**形容词、副词、动词和名词**中，这一步是为了筛选出有潜力的“候选词”，去除无关干扰。

---

3. 第二部分：中间核心——双重算法并行（双保险）

讲解重点： 这是整个框架最“硬核”的部分。你需要解释图中的上下两条路径是在做 **互补** 。

- **图示对应：** 中间的大虚线框。
- **逻辑解释：**
  - 我们将“候选词”与已有的“种子词典”（基准情感词）进行对比。为了保证准确性，我们采用了两种不同的计算维度：
  - **维度一（上方路径）：基于统计的共现概率（PMI）。**
    - 公式：$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$。
    - **通俗解释：** 计算候选词与种子词同时出现的概率。如果一个词经常和“高兴”一起出现，那它大概率也是正面的。这里通过阈值 H1 进行初步判断 ^2^。
  - **维度二（下方路径）：基于语义的相似度（余弦距离）。**
    - **通俗解释：** 利用词向量（Word2Vec等）计算语义空间的距离。如果一个词在向量空间里离“悲伤”很近，那它大概率是负面的。这里通过阈值 H2 进行判断 ^3^。

---

4. 第三部分：集成与迭代（滚雪球）

讲解重点： 这是该框架的 **亮点** 。它不是一次性算完，而是循环进化的。

- **图示对应：** 中间右侧的“集成规则”菱形框 → “添加当前词” → 回到起点的箭头。
- **逻辑解释：**
  - **集成规则：** 单一方法可能有偏差，所以我们结合 PMI 和 余弦距离 的结果进行 **集成判定** 。只有双重校验通过的词，才会被认定为新发现的情感词。
  - **迭代扩展（关键）：** 这一步是点睛之笔。大家请看步骤 4.3 和 5—— **新发现的情感词，会被立即加入到“种子词典”中** 。
  - **意义：** 这意味着我们的“标尺”在不断变长。下一轮计算时，我们用新加入的词去寻找更多的词。这是一个**自举（Bootstrapping）**的过程，直到没有新词产生为止。

---

5. 总结升华

话术：

> “**最终，通过这种 统计与语义相结合 、****循环迭代**的方式，我们不仅解决了人工构建词典成本高的问题，还能够有效捕捉到社交媒体中不断涌现的新造情感词，构建出高质量的专用情感词典（如右侧输出所示）^4^。”