## 1. 假期任务简单概述

*简而言之就是先用最低的成本设计一个整套的产品，后面确定可行后逐渐复杂化，采用更复杂的模型和更大规模的系统。*

1. **需要训练的模型**：股价预测模型
2. **需要收集/爬取的数据：**
    - 2020～2025 年的历史行情数据，但是着重研究单支股票。（选取沪深 300？）
    - 响应的时间的财经新闻内容
3. **不需要训练的模型：** FinBERT 或者 SnowNLP，只需要接入 API 就行
4. **多模态融合：** 手动加权新闻情感分数和股价预测的模型, 例如

```
final_prediction = 0.7 * price_prediction + 0.3 * sentiment_score
```

注意，我们只需要突出核心创新点和确保预测的准确性就行：股价数据是核心创新点（如何结合新闻和时序数据），所以需要自定义训练。

---

## 2. 具体细化：

1. **数据**
    - 数据来源
        - 股价数据（行情数据）：tushare Pro，akshare，Yahoo finance
        - 新闻文本：新浪财经；示例

```
https://search.sina.com.cn/?q=贵州茅台
&range=title&c=news&sort=time
```

        - 舆情：
            - 东方财富网：茅台股吧讨论（过滤噪声）
            - 财联社：快讯新闻
            - 微博：热搜话题（关于茅台的）
        - 公开数据集：
            - kaggle 等
    - 研究对象初筛：优先选取中国市场
        - 贵州茅台（A 股，600519.SH）
        - 苹果（美股，AAPL）
            - 数据完整（Tushare/Yahoo免费获取）
            - 新闻舆情丰富（财报、行业政策、管理层变动等）

        - 价格趋势清晰（适合验证时序模型）

- 研究对象后续扩展：
    - A 股多因子预测
        - 沪深 300 成分股（300 只）
        - 2020～2025 年日频数据+2024 年分钟级数据
        - 字段：OHLCV + 换手率 + 北向资金净流入

- 数据量
    - 股价：2020～2025，选择日频
        - 字段：OHLCV + 换手率 + 北向资金净流入
    - 新闻：2020～2025, 规范：时间戳+新闻标题+主要内容（去广告等）

| 新闻类型 | 最低数量 | 时间范围 |
| --- | --- | --- |
| 财报/公告 | 25～30% | 近3年（2021-2024） |
| 行业政策 | 20～25% | 近5年（2019-2024） |
| 管理层动态 | 15～20% | 近5年 |
| 社会舆情 | 10～20% | 近3年 |

    总计：收集 500 条以上高质量数据

- 数据清洗
    - 行情数据
        - 缺失值处理，时间戳标准化
    - 新闻文本
        - 建立新闻舆情分析的停用词库
        - 数据标注，数据标注的规范制定，数据标注的算法自动化处理
        - 使用 FinBERT 计算情感得分

- 数据标注规范、数据标注
    - 要制定一套规范的合理的数据标注规范
    - 编写 Python 脚本进行自动化数据标注 示例：

```
{
  "text": "茅台提价10%引发经销商囤货",
  "label": "positive",  # 正面/中性/负面
  "intensity": 0.8      # 强度（0-1）
}
```

2. **模型**

- 时序模型选取（股价预测）

- 时序模型训练
- FinBERT + SnowNLP
  - 用SnowNLP快速过滤无关新闻（如情感分数在0.4-0.6之间的中性新闻可跳过）。
  - 对剩余新闻用FinBERT深度分析（需微调中文金融数据 ）。
    - 注意这里的微调需要 GPU 资源和大量高质量文本
- 与FinBERT 预训练模型加权输出

3. **web 简单搭建**（见后文）

- 前端工具
- 后端工具
- 部署工具

4. **技术栈和硬件需要**

- 首先满足轻量化和低成本的需求：
  - 存储
    - 原始数据：Google Drive（15G）
    - 清洗后的结构化数据：Github 私有仓库（LFS）+新增的数据更新
    - 数据量：
      - 行情数据，10MB
      - 新闻数据：70～170MB
      - 模型训练数据：500MB的 FinBERT 数据和 5MB 的时序数据
  - 运行
    - 开始的数据先存储在 Google Drive 上面，清洗完了用 GitHub 私有存储
    - 用 kaggle Notebook 直接挂载 Google Drive, 每周 30 小时的免费 GPU训练 transformers 模型
  - 部署
    - 模型权值存储在 Hugging Face Hub, 前端数据存储在Superbase 免费版