3.15 模型部分大纲

一、通过网络评价等数据，构建一个情感指数

- 核心数据

数据类型	来源平台	采集字段示例
短文本评论	豆瓣电影、微博超话	评论内容、发布时间、点赞数、回复数
长文本评论	知乎影评、猫眼专业影评	评论文本、评分（1-5星）、用户等级
结构化评分	猫眼、淘票票	每日评分、评分人数、各星级占比
衍生行为数据	短视频平台、搜索指数	电影相关视频播放量、弹幕情感倾向（需NLP解析）

ADVICE: 怎么数据清洗（eg 水军）

- 情感计算模型
  - 传统统计/NLP 方法（短评）：使用TF-IDF加权的扩展情感词典方法
  - 深度学习模型（长评）：基于RoBERTa-wwm-ext的微调 / 调用大模型 api
    - (**ADVICE1**: 不要直接打分，不然和豆瓣上的五个星星没啥区别了。 我的想法：让模型根据评论内容，给情节、主题……多方面，每个方面打分)
    - (**ADVICE2**: 存在评论和打分不匹配的情况，比如评论不好但是打分高，针对这种情况我们可以分析是否存在 discrepency or domination)
  - 两个模型动态加权
    - (**ADVICE**: 怎么加权？)
  - 引入可信度/权重（可能可以用清风的方法），区分水军、普通人、专业影评 & 不同平台数据来源
    - (**ADVICE**: 怎么得到这个置信度？)
  - 引入时间衰减

- 指数合成

$$SEI_t = \frac{1}{1 + e^{-(0.5S_t + 0.3V_t + 0.2I_t)}} \times \log(1 + N_t)$$

  - S_t：当日情感得分均值
  - V_t：情感方差（反映舆论分歧）
  - I_t：影响力加权分（点赞数×用户可信度）
  - N_t：当日评论总数

- 统计检验
  - 建模过程中：F1、MSE、R^2 这些指标
    - **Question**: 没有 ground truth 怎么检验?
  - 格兰杰因果检验：验证情感指数对票房的领先性
  - 可视化：叠加情感指数和票房走势

二、将情感指数和其他结构化数据一起，构建一个机器学习模型，用于预测票房走势

- 核心数据
  - 电影属性
    - **基本特征**：类型、导演、主演、制作公司、预算、片长、续集/IP改编、分级（如PG-13）。
    - **历史表现**：导演/演员过往作品的票房、评分（如用加权平均或衰减因子处理）。
    - **上映信息**：上映日期（节假日/周末）、发行地区、影院数量、排片率（首周及后续）
  - 情感指数
- 模型构建
  - XGboost / 随机森林：更注重结构化特征
  - LSTM：更注重时间序列
    - **ADVICE**: LSTM 太老了，transformer + positional encoding / mask
    - **posible models**: Informer, Autoformer, LTSF-Linear
    - 但是我们不一定有这么多数据?

    - MoE 结合两个模型
- 统计检验
    - F1、MSE、R^2 .....
    - SHAP
    - 同时也构建一个 arima 模型，并将我们的模型和这个传统模型进行比较
    - 案例分析：用模型分析有代表性的电影，分析大误差电影的可能原因