股票趋势预测与新闻情绪关联分析

一、研究背景及意义

金融市场是一个复杂且动态变化的系统，受到多种因素的影响。传统的股票价格预测模型多基于历史股价和成交量等数据，但随着信息传播速度的加快和互联网技术的发展，新闻报道、社交媒体等成为影响投资者决策的重要因素。个人投资者在股票投资中常面临信息过载问题，难以从海量市场评论中提取有效情绪信号。本项目旨在通过构建一个结合了股票历史数据与相关新闻情绪分析的预测模型，结合用户自己的风险承受能力，来做出股票的portfolio推荐。

二、小组成员

组长：肖景泽

负责整体项目的规划、监督以及模型设计与算法实现。

组员1：刘禹阳

主要负责数据收集与预处理工作，包括但不限于从Yahoo Finance抓取股票历史数据，从新浪新闻、腾讯新闻等平台抓取相关新闻文本数据，并进行清洗和格式转换。

组员2：陈浩

- 主要负责结果分析和报告撰写工作，对模型预测的结果进行深入分析，评估模型性能，并根据分析结果提出改进意见。

三、数据来源和获取方式

* **股票历史数据**：将使用Python脚本从Yahoo Finance等公开金融数据平台自动下载所需时间段内的股票历史价格、成交量等数据。为确保数据的准确性和完整性，我们还将对下载的数据进行验证和补充。
* **新闻文本数据**：采用网络爬虫技术从主流新闻网站（如新浪新闻、东方财富）抓取与目标股票相关的新闻文本。针对不同网站的数据结构特点，编写相应的爬虫程序，并对获取的数据进行去重、分词、停用词过滤等预处理操作。

四、问题建模与描述

本项目属于回归问题，核心在于通过分析股票历史价格、成交量以及相关新闻的情绪倾向，预测未来一段时间内股票价格的变化趋势，并向用户做出股票的portfolio推荐。我们将综合运用时间序列分析方法和自然语言处理技术，提取并融合多源异构数据特征，以提高预测模型的准确性。

五、解决途径与模型

**长短期记忆网络（LSTM）**：考虑到股票价格具有显著的时间序列特性，采用LSTM模型能够有效捕捉长期依赖性和动态变化规律，特别适用于处理时序数据。

**情感分析模型（如BERT）**：利用深度学习技术对新闻文本进行情感分析，将其转化为数值型特征，以便与股票价格数据相结合，共同作为输入特征参与模型训练。

六、进一步改进的思路

1. 引入更多外部因素，如宏观经济指标、行业动态等，进行多维度特征融合，增强模型的解释力和预测能力。
2. 探索更高级的自然语言处理技术，如词嵌入和注意力机制，以提高文本特征提取的准确性，进而优化模型性能。
3. 预期成果
4. 构建一个能够有效结合股票历史数据与相关新闻情绪分析的股票趋势预测模型，能够在一定程度上反映市场情绪对股票价格的影响。
5. 通过上述研究，我们期望不仅能够提升股票价格预测的准确性，同时也希望能够为投资者提供更加科学的投资决策依据，来做出具体的股票的portfolio推荐。