实验思路主要参照论文《ChatGPT, Stock Market Predictability and Links to the Macroeconomy》

代码在<https://github.com/Atticsu/NLP_based_marktet_risk_predict>

介绍在Readme文件中

主要步骤为：

1. 获取数据：编写爬虫，爬取新浪财经快讯（只保留有“重要”标识的 防止噪声）的全样本历史数据

<https://finance.sina.com.cn/7x24/>

1. 情感分析：

1）目前主流情感分析很多依赖大模型进行，而大模型收费较贵（快讯数据量较大，约20万条，调用API开销可能较贵）因此采用调用在金融语料上特训过的FinBERT模型进行情感分析作为替代

2）FinBERT模型的原版模型是基于英文的，在中文表现一般；因此采取了两种方式进行，分别为：a)调用港科发布的基于中文微调后的finbert\_tone\_chinese模型;b)调用Helsinki-NLP/opus-mt-zh-en将原始数据翻译为英文，再调用原版FinBERT，最终两个的效果都输出去比较

模型输出为：该快讯情绪为Positive/Neutral/Negative的概率值，当且仅当某项情感置信度超过**95%**时确认情感类别

1. 指标构建和评价：指标参照论文方式构建，较为简单，分别为pov\_ratio和neg\_ratio两个

Pov\_ratio: 正面情绪的新闻占总新闻的比重（月度）

Neg\_ratio：负面情绪的新闻占总新闻的比重（月度）

回归方程：



其中，为因变量，t+h为覆盖的时间窗口，h分别采取0（当月）,1,3,6,9,12，分别检验了

上证指数（sh\_index），中证1000（cn\_1000），中证500（cn\_500）,科创50（tech\_50）的收益率，标准差（成分股，采取最新成分股数据，由于数据可得性问题，未根据历史调整），夏普比率（收益/波动）

Ratio为检测的指标，pov和neg分别检测

结果在Excel中，此处不再重复

Excel中因变量命名：

{xxx}\_lag{h} 为滞后h窗口的累计收益率

monthly\_return monthly\_sharp 1\_month\_std: 为当期收益率/夏普/标准差

index\_name对应为：上证指数（sh\_index），中证1000（cn\_1000），中证500（cn\_500）,科创50（tech\_50）