接下来介绍项目的 proposal
首先是**研究的背景与动机**
中国股市的投资者通过在股票论坛上发布帖子来表达对不同行业股票的关注。这种关注度可以反映特定行业的市场表现。目前，衡量行业关注度的研究在中国股市中还相对缺乏。

在**动机**方面
过往研究表明，投资者关注度对股票价格具有预测能力，并与股票收益和交易量相关。
其次，人工智能技术在金融分析中变得更加重要，需要探索更先进的 NLP 自然语言处理方法来处理非结构化文本数据，尤其是在中国股市的背景下。

具体到我的**研究问题**
核心问题是研究投资者关注度对股票收益和交易量的影响
具体到以下四个问题：
1.如何定义和量化特定行业的投资者关注度？
2.在线股票论坛中哪些文本数据可以支持研究？
3.可以采用哪些方法来分析文本数据？
4.投资者关注度指标与交易数据（如股票价格涨跌和交易量）之间的相关性如何？

对于**研究方法**
首先我们需要对关键变量进行定义：
我们用行业关注度指数（Industry Attention Index，简称 IAI）作为量化指标，衡量特定行业投资者关注度。该指数基于特定行业最具代表性股票的活跃帖子数量来计算。

计算公式： $IAI_i = \sum_{k=1}^{N} V_{i,k}$ 其中，N 是所选股票的数量，这些股票最大程度上代表了该行业，$V_{i,k}$ 是在单个时间周期内（例如过去30天）股票k的帖子数量。

这个 IAI 表示对于行业 i, 其中具有代表性的 N 只个股 k 在单个时间周期内的帖子数量总和。

然后定义 sentimentIAI 是基于投资者帖子的情绪分类结果计算出的关注度指标。把帖子分为正面情绪和负面情绪两类，并分别计算其帖子数量。

计算公式： $sentimentIAI_i = \frac{\sum_{k=1}^{N} V_{i,k}^{positive}}{\sum_{k=1}^{N} V_{i,k}}$ 其中，$V_{i,k}^{positive}$ 是股票k在行业i中正面情绪帖子的数量，$V_{i,k}$ 是总帖子数量。情绪IAI的值越高，表示正面情绪帖子占总帖子的比例越大，反映出投资者对该行业的乐观情绪。
就是用 positive 的 IAI 除以总的 IAI，这个指标值越大，反映出投资者对于该行业的乐观情绪。

在**数据收集**方面
文本数据考虑从东方财富网（也就是股吧），还有雪球等网站收集。我们需要更关注个股论坛的帖子，因为它们会表达强烈的买入和卖出偏好。
市场交易数据包括每日开盘、收盘、交易量和金额等指标。

然后是 **sentimentIAI 的分类任务**：
设计深度学习情绪分析模型来对股票帖子的情绪倾向进行分类。
模型结构包括（embedding 层、LSTM 层、CNN 层、attention 层和输出层）
embedding 层：　使用预训练的 word 2 vector 模型构建文本的词向量。

LSTM 层： 使用长短期记忆网络来捕捉上下文信息。
CNN 层： 在 LSTM 的输出上执行卷积操作，提取文本特征。
attention 层： 通过赋予不同的概率权重并乘以 CNN 输出向量，然后通过 softmax 函数计算权重矩阵。
输出层： 通过 sigmoid 函数映射结果，进行情绪分类。

最后是**回归分析**:
尝试使用各类型的回归分析方法来检验 IAI 和 sentimentIAI 与股票超额收益、交易量、收益波动率的因果关系。预测两种 IAI 对超额收益、交易量、波动率的回归结果应该显示出显著影响。