# 运用新闻和公司信息预测医疗行业股市波动

钱丽玥
SIST
ShanghaiTech University
Shanghai, China
qianly@shanghaitech.edu.cn

沈昭正
SIST
ShanghaiTech University
Shanghai, China
shenzhzh@shanghaitech.edu.cn

井文可
SIST
ShanghaiTech University
Shanghai, China
jingwk@shanghaitech.edu.cn

## 1 BACKGROUND

### 1.1 Introduction

金融科技高速发展的今天，股票市场已经成为了一个国家金融体系中非常重要的一部分。长期以来，由于股票价格具有高度波动性，使得股票价格的预测成了分析师和研究人员关注的话题。实际上股票价格受到众多因素的影响，包括领导层的变化、国家的货币政策，行业的景气状况和投资者的情绪等等。其中投资者对财经新闻的反应是一个至关重要的影响因素，在自媒体高速发展的今天这一点更加明显。财经新闻以一种公开的信息披露方式，对我国股市产生了非常重要的影响，其中包括政策扶持类财经新闻、兼收并购类财经新闻、再融资类财经新闻、盈利类财经新闻、违规处罚类财经新闻。尽管它们对股市的影响不尽相同，但在我们的课程作业中将不区分新闻本身的属性，而只关注新闻对于股市的利好利空影响。现有的大量研究已经将股票相关的新闻信息作为预测股票价格的重要依据，结果表明股票价格的变动与新闻文章的发布之间存在很强的相关性，已尝试使用支持向量机，朴素贝叶斯回归和深度学习等算法在各个级别上进行了几种分析研究。

同时近些年来，随着人民生活水平的提高和对医疗保健需求的不断增长，我国医药行业一直保持着较快的增长速度，在国家政府对行业的大力扶持下，医药行业逐渐成为国民经济中发展最快的行业之一。2020 年世界范围内新冠肺炎疫情的爆发，进一步提高了国家乃至世界对医药行业的重视程度，并且加大了在资金、技术、政策方面的支持和帮助。此外，在人口老龄化以及中国城镇化加速的背景下，居民收入普遍增长加上人们愈发重视医疗保健，医保体系进一步健全，这些都将刺激并提升中国医药消费需求，推动整个医药行业进行新的变革并形成新格局。因此在我们的课题研究中，将着重探讨医药行业的新闻对股票价格的影响。

### 1.2 Related Work

在《基于社交情感数据挖掘的股票市场预测研究》[1]中，梁士利等人以结合粒子群阈值优化改进的贝叶斯算法为基础，以股吧等金融背景的语料库为关联规则挖掘的对象，利用 jieba 分词工具来统计每个词出现的频率，人工筛选标注出不同情感极性的词语，建立专属的金融情感词典。并利用自然语言分析技术，研究句子级别中新词识别与句法 中否定词对情感分析的重要性，对每条评论的情感倾向进行自动识别，从而完成对股票市场的分析。

在《股票预测：一种基于新闻特征抽取和循环神经网络的方法》[2]一文中，张泽亚等人提出了一种基于单词点互信息的新闻特征抽取方法，将该特征运用到股票价格的预测中，并提出了一种基于循环神经网络的股票预测模型。具体而言，股票预测问题的输入包含两种不同类型的输入，即新闻文本数据和股票价格数值数据。但是由于文本数据不能够直接作为分类器输入，所以作者首先选取利好关键词和利空关键词的种子集合，然后使用最优化方法选出利好与利空的标准关键词集合，并用标准关键词集合计算其他单词的利好极性。之后，为了分析一则新闻是利好新闻还是利空新闻，作者运用 One-Hot 方法，根据已有的已经定义好所有单词利好极性的最优标准集，来判断新闻的属性。最后考虑到股票价格在短期内有着一定的时序相关性以及新闻事件对于股票价格的影响具有持续性，作者提出循环神经网络模型。从实验结果来看发现相对于基于价格特征的 SVM 分类器，张泽亚等人提出的方法在股票涨跌预测方面能有超过 5%的提升。

在论文《Stock Price Prediction Using News Sentiment Analysis》[3]中，Saloni Mohan 等人利用 NLTK 库来分析新闻文本的情感极性，包括积极和消极情感 (N, P)。每个新闻文本中的极性标识为最大绝对极性，即 Polarity = (+/−)max(abs(N, P))，而一个公司的最终极性等

于其对应的所有文本的平均极性。他们根据当前的极性$P_t$和时间 t 之前的股票价格$Price_{t-1}$来训练 RNN 模型，也就是说根据$(Price_{t-1}, P_t), (Price_{t-2}, P_{t-1}), \dots, (Price_{t-m}, P_{t-m+1})$，可以预测出 $Price_t$的股价。此外，论文还采用了另一种方法，与前面的方法不同，不是分析新闻文本的情感极性，并将整个文本与股市价格一起作为 RNN 的输入。

## 2 AIMS AND OBJECTIVES

新闻作为信息传播的一大媒体，它对股市波动有着强烈的影响。例如论文《Textual analysis of stock market prediction using breaking financial news The AZFin text system》[4]中就有加入新闻与不加入新闻输入的模型对比，说明了新闻的重要性。一则新闻的发布，对新闻所关注的公司有着或轻或重的影响，影响程度与新闻的发布时间、发布平台、正文内容、点击量、标题等有关。我们主要关注标题的发布时间、内容对股市波动的影响。

同时，不同新闻的点击量与新闻发布的平台相关，点击量越多的新闻影响力也越大，对股票波动的程度影响也越大。所以，我们打算同时将新闻发布的平台这一因素纳入对股票波动预测的分析当中。

此外，因为不同公司可能存在着一些联系。例如，同一个人在不同公司都有重要职务 ，那么某一家股票的涨跌也可能影响与其相关的公司波动。如果可能的话，我们也想利用这些信息，构建图谱，来帮助预测医疗行业的股票。总的来说，我们的目标包括下面几个方向：

1. 通过新闻文本的关键信息提取对所选股票的涨跌幅预测
2. 考虑新闻点击量对股票波动幅度的影响
3. 利用公司信息和联系找到股票之间的波动关系

## 3 PROPOSED METHODOLOGY

我们选取了东方财富和新浪财经 2020 年 5 月至 12 月的财经新闻数据，包括 14 万余条新闻的发布时间、来源、标题、正文。新闻的信息如图一。从 14 万余条新闻中筛选出与所选的医药行业的股票名称匹配和相关的新闻，过滤掉无关的新闻。再对筛选后的新闻文本利用 jieba 对新闻文本进行分词，创建出两类情感词典：积极词典（上涨、反弹等）和消极词典（下跌、跌停等）。对于每一篇新闻，我们统计出内容的积极和消极程度，从而帮助之后的训练。在提取出新闻的积极消极程度特征之后，在同花顺金融服务网爬取对应时间内每日收盘价，将这些数据送入 RNN 进行训练，预测股票的涨跌情况及概率。

| tid | Datetime | Source | Titile | Text |
|---|---|---|---|---|
| c81b09334d2f38d4e831f549b51998db | 2020-08-27 22:22:00 | 东方财富网 | 核心产品"集采"失标后 华东医药业绩保持增长 | 华东医药(000963)核心产品阿卡波糖片在"集采"的失标，曾让市场担忧公司今年的业绩。不过，实际情况却是公司上半年业绩依旧保持逆势增长…… |

**Figure 1: 截取了一条医药行业的新闻数据，其信息包括新闻 id、发布时间、来源、标题、正文。**

## 4 FEASIBILITY

### 4.1 Data Feasibility Analysis

我们选取的财经新闻数据是从东方财富和新浪财经中提取出来的，数据本身的可信度较强，不会存在较大噪声。同时，5 月至 12 月的财经新闻数据有 14 万余条，整体数量较多，我们简单观察了一下新闻数据，其中的关于医药行业的新闻数量也不少，基本可以按照标题判断新闻在分析哪只股票，可能对哪些股票有影响。同时，我们可以从同花顺金融服务网爬取对应时间的股票信息，信息比较全面，能够获取和利用。

### 4.2 Method Feasibility Analysis

我们将要采用的是 RNN 方法，输入前一天的股票收盘价以及当天的新闻特征提取结果来预测当天的股票收盘价格涨跌幅。在 Saloni Mohan 等人的研究中[3]，大多数 RNN 方法的性能都非常好。平均绝对百分比误差（MAPE）是预测模型的预测准确性的常用度量，他们用该度量对不同方法的效果进行评测。其中 RNN 方法的 MAPE 值在 2.03 和 2.17 之间，然而 ARIMA 和 Facebook Prophet 等经典时间序列模型的得分要低得多，介于 7.39 和 7.98 之间。使用财经新闻文章作为输入内容的模型表现很好，而仅根据历史股价预测未来股价的模型会导致较高的百分比误差。所以该方法具有一定的可行性。

### 4.3 Computing Ability

由于我们需要通过 RNN 进行训练，可能需要有比较强的运算能力的电脑。

## 5 TIMETABLE AND WORKLOAD DIVISION

| 任务列表 | 钱丽玥 | 沈昭正 | 井文可 |
|---|---|---|---|
| 阅读文献 | ✅ | ✅ | ✅ |
| 讨论实现方法 | ✅ | ✅ | ✅ |
| 爬取股票数据及预处理 | ☑ | | |
| 筛选新闻数据及预处理 | | ☑ | ☑ |
| 训练 | ☑ | ☑ | |
| 测试 | | | ☑ |
| 分析实验结果 | ☑ | ☑ | ☑ |
| 完成报告、ppt | ✅ | ✅ | ✅ |
| presentation | ☑ | ☑ | ☑ |

**Figure 2: 任务列表的分工和完成情况。绿色对勾表示完成了以及正在完成的内容；灰色对勾表示计划完成的内容。**

## REFERENCES

[1] 梁士利,陈翌昕,陈培培,孙丽敏.基于社交情感数据挖掘的股票市场预测研究[J].东北师大学报(自然科学版),2020,52(03):105-110.

[2] 张泽亚,黄丽明,陈翀,闫宏飞.基于新闻特征抽取和循环神经网络的股票预测方法[J].文献与数据学报,2020,2(01):45-56.

[3] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[4] Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst. 27, 2, Article 12 (February 2009), 19 pages. DOI:https://doi.org/10.1145/1462198.1462204.