



北京大学

硕士研究生学位论文

题目： 基于新闻特征与 LSTM
模型的股票预测方法

姓 名： 张泽亚

学 号： 1401214271

院 系： 信息科学技术学院

专 业： 计算机系统结构

研究方向： 搜索引擎与网络挖掘

导师姓名： 闫宏飞 副教授

二〇一七 年 六 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

股票市场是资本市场的重要组成部分，股市的波动与市场经济息息相关。股票预测问题伴随股票市场的建立而一直存在。人们希望找到有效预测股票走势的方法，不论是在计算机领域还是在金融领域，也不论是从理论角度还是从实证角度。然而股票预测问题极具挑战。一方面因为股票预测需要运用经济学、计算机学与数学等多学科知识，另一方面因为股票市场受随机事件的影响而波动不稳定。近十几年，随着计算机技术在大数据与人工智能方面的突破进展，人们有望在股票预测方面取得新的进展。

本文研究基于新闻语料的股票预测方法。该预测方法首先面向股票预测问题，从股票相关新闻中抽取出特征，然后将新闻特征融入 **LSTM** 预测模型，对股票涨跌做出预测。在特征抽取方面，本文提出了一种基于单词利好极性的新闻特征抽取方法。该方法先依据经验选取一些能够代表利好新闻与利空新闻的单词，基于最优化方法计算出所有单词的利好极性。之后基于新闻中单词的利好极性，构造出新闻的特征向量。预测模型方面，本文提出了一种融入新闻特征的 **LSTM** 预测模型，将价格的时序性以及新闻影响的持续性反映在模型中。为了检验预测方法的效果，本文在中国股票市场与美国股票市场分别进行了实验。在股票涨跌预测实验中，相比基于价格特征的 **SVM** 预测方法，本文提出的方法在股票涨跌预测问题上有 2% 至 3% 的准确率提升。在股票模拟交易实验中，本文提出的新闻特征和预测模型有助于提高回报率，但是该算法离实际应用还有一定距离。

关键词：股票预测，特征抽取，**LSTM**

A Stock Prediction Method based on News Features and the LSTM Model

Zeya Zhang (Computer Architecture)

Directed by A.P. Hongfei Yan

ABSTRACT

Stock market is an important part of the capital market, the stock market volatility and economy is closely related. The problem of stock prediction has been around with the establishment of the stock market. People want to find a way to effectively predict the stock trend, whether in the computer or financial field, both from a theoretical or an empirical point of view. However, stock forecasting is a challenge. On one hand it is because the stock forecast requires the multidisciplinary knowledge of economics, computer science and mathematics, on the other hand it is because the stock market is influenced by random events. Over the past decade, with the progress of big data and artificial intelligence, people are expected to make new progress in stock prediction.

This paper studies the stock prediction method based on news corpus. The prediction method first extracts the features from stock-related news, and then integrates the news features into a LSTM model to predict the stock trends. In the aspect of feature extraction, this paper proposes a method of extracting news feature based on the word polarity. The method first select some words that can represent good news and bad news by experience, and then calculate the positive polarity of all words based on an optimization method. After that, construct the news features based on word polarity. In the aspect of model, this paper proposes a LSTM prediction model which incorporates news features. This model reflects the timeliness of price and the persistence of news influence. To test the effect of prediction methods, this paper carries on the experiment both in the Chinese and American stock market. In the experiment, the method proposed in this paper has a 2% to 3% improvement in the problem of stock price fluctuation prediction compared to the price-based SVM method. In the trading simulation experiment, the news feature and prediction model proposed can improve the rate of return, but the algorithm is still a distance from the practical application.

KEY WORDS: Stock Prediction, Feature Extraction, LSTM

目录

摘要	I
ABSTRACT	III
目录	V
图形列表.....	VII
表格列表.....	IX
第一章 引言.....	1
第二章 相关工作.....	3
2.1 股票价格预测	3
2.2 情感分析	4
2.3 长短时记忆网络 (LSTM)	5
2.4 小结	6
第三章 问题定义.....	7
3.1 股票市场	7
3.2 股票相关信息	8
3.3 股票涨跌预测问题	8
3.4 小结	9
第四章 算法详述.....	11
4.1 算法综述	11
4.2 特征抽取	12
4.2.1 股票价格特征构造	12
4.2.2 单词的利好极性.....	12
4.2.3 股票新闻特征构造.....	16
4.3 预测模型	17
4.3.1 LSTM 层	17
4.3.2 LSTM 预测模型	20
4.3.3 模型参数学习	22
4.4 小结	23
第五章 实验.....	25
5.1 数据集	25
5.1.1 数据采集与处理	25
5.1.2 数据集统计	29

5.2 实验结果.....	29
5.2.1 单词利好极性.....	29
5.2.2 股票预测.....	30
5.2.3 投资回报率.....	32
5.3 小结.....	33
第六章 结论与未来工作.....	35
6.1 结论.....	35
6.2 未来工作.....	35
参考文献.....	37
附录 A 符号表.....	41
附录 B 硕士期间科研成果.....	43
附录 C 股票数据集.....	45
附录 D 部分公式推导.....	47
致谢.....	51
北京大学学位论文原创性声明和使用授权说明.....	53

图形列表

图 4.1	股票预测算法框架	11
图 4.2	Memory Block of LSTM	18
图 4.3	LSTM 预测模型	20
图 5.1	SeekingAlpha 网站中股票的数据页面	26
图 5.2	SeekingAlpha 网站中股票的相关新闻页面	27
图 5.3	新闻页面部分 HTML 代码	28

表格列表

表 4.1	LSTM 层伪代码	19
表 4.2	LSTM 模型预测过程伪代码	21
表 5.1	股票相关数据来源	25
表 5.2	实验使用的中国股票相关信息统计	29
表 5.3	实验使用的美国股票相关信息统计	29
表 5.4	极性值处于两端的单词	30
表 5.5	国内股价涨跌预测准确率比较	30
表 5.6	美股股价涨跌预测准确率比较	31
表 5.7	34 只美股上模拟交易实验结果	33

第一章 引言

股票市场是上市公司筹集资金的主要途径之一，是资本市场的重要组成部分。股票市场被称为“经济的晴雨表”，它的波动与市场经济的兴衰息息相关。

股票对于上市公司来说是一种融资方式。上市公司通过在股票市场发行股票，筹集社会上的闲散资金以解决资金暂时不足的困难。股票对于个人来说是一种投资理财产品。持有合适的股票能够让个人资产升值，例如伯克希尔·哈撒韦公司¹的股票，股价由 2000 年的 5 万美元一路稳定增长至 2017 年的 25 万美元。股票对于许多基金公司与证券投资公司来说，是一种主营业务。随着计算机科学技术的发展，越来越多的公司开始将大数据与人工智能技术运用于金融投资领域。例如有新闻报道²，2000 年高盛³位于纽约的股票现金交易部门有 600 多位交易员，而到 2017 年初，只剩下 2 名交易员，其余的工作全部由计算机包办。李开复也曾经说：“高盛这样的金融巨头，以及其他大型对冲基金，都正在转向由人工智能驱动的系统，以预测市场趋势，从而做出更好的交易决定。”

股票预测问题随着股票市场的建立而一直存在。股票价格可以预测么？Fama 于 1965 年提出了有效市场假说^[1] (Efficient Market Hypothesis)，他认为，股票市场是“有效信息”市场，即股票价格充分反映了已经发生的事件，以及那些尚未发生但是人们预期会发生的事件对股票价格的影响。这一理论成为后续股票预测工作的基础和依据。Fama 按照市场信息量的大小，以及市场对信息的反应速度，把有效市场分成三个层次：弱势有效、半强势有效和强势有效。中国股票市场属于弱势有效市场，即当前股票价格能反映历史交易信息，却不能及时体现公开信息的作用。美国市场属于半强势有效市场，即股票价格能及时地综合反映历史交易信息和市场公开信息。

今天，随着计算机科学技术的进步，股票市场正在揭开神秘的面纱。首先，大数据与云计算的兴起，让人们在处理与分析海量股票交易数据时变得游刃有余。在基金公司和证券投资公司中，量化分析正在从人工逐渐走向智能。其次，深度学习在自然语言处理领域取得突破，计算机收集和分析股票市场公开信息的效果越来越好。此外，机器学习技术不断进步，计算机越来越“智能”，计算机很有可能是解决股票预测问题的新钥匙。2016 年 3 月，阿尔法围棋 (AlphaGo) 凭借深度学习技术，以 4:1 的总比分战胜世界围棋冠军李世石。既然计算机能够在围棋这一古老而深奥的智力游戏中战胜人类顶尖棋手，那么有理由相信，计算机也有

¹ Berkshire Hathaway Inc. 美国一家世界著名的保险和多元化投资集团，由沃伦·巴菲特创建。

² <http://business.sohu.com/20170215/n480761931.shtml>

³ 美国高盛集团 (Goldman Sachs)，一家国际领先的投资银行。

能力在股票市场超越人类证券分析师。

综上所述，一是股票市场在经济中有着重要的地位，二是股票对于个人和公司都有重要的价值，三是股票预测在今天正愈加具备可行性。因此，股票预测方面的研究变得很有意义。

本文的余下部分组织如下：

第二章介绍了股票预测领域重要的相关工作；

第三章给出了股票预测问题的形式化定义；

第四章详细论述了本文提出的股票预测方法；

第五章测试了算法在真实股票数据上的效果；

第六章对本文总结归纳，并展望了可能的未来工作。

第二章 相关工作

2.1 股票价格预测

股票价格受到诸多因素的影响，因此股票价格预测问题极具挑战性。在宏观层面，股票市场是市场经济的重要组成部分，受到财政政策和货币政策的影响，股票价格的不确定性增加。在微观层面，市场上存在内幕交易等违规活动，这加剧了股票市场的信息不对称，提高了股价预测的难度。正如 Fama 所说，许多股票市场处于弱势有效，市场的公开信息不能及时的作用到股票价格中，而是需要一段反映时间。这种市场信息的延迟性扩大了信息在股价上的作用时间，给预测股价设置了障碍。

考虑到市场信息对于股价的作用，许多相关工作^[2-4]都把与股票相关的新闻作为预测股票价格的重要依据。然而，新闻文本本身难以直接用于预测股价，人们需要从新闻中提取出有效的市场信息才能发挥新闻的价值。这本质上是一个自然语言处理问题，但因为特征抽取是针对金融市场的，所以在自然语言处理时需要结合这一领域的相关知识。这无疑给股票预测问题带来了新的困难。

学术界对股票预测问题有着浓厚的研究兴趣。

1952 年，美国经济学家 Markowitz 发表学术论文《资产选择：有效的多样化投资》^[5]。论文中首次运用资产组合回报的均值与方差这两个量化投资指标。Markowitz 从数学上形式化定义了投资者偏好，用数学方式解释了投资分散化原理，并系统阐述了资产组合与投资选择问题。这一成果标志着现在投资组合理论的开端。

此后，有的学者通过分析大众媒体和社交媒体上群众的言论情感，对股票市场的整体走势（道琼斯工业指数、上证指数等）做出预测^[6-9]。Johan Bollen 等在论文^[10]中利用推特（Twitter）网站上的推文内容对道琼斯工业指数进行预测。他们使用情感分析工具分析推特上每日的大众情绪，然后将这些情感变量作为特征加入到预测模型中，预测股票市场的涨跌。

更多的学者不局限于预测股市的整体走势，而是深入研究每只股票的价格波动。有的学者通过统计学习方法预测股价走势，例如 GPC Fung 等在论文^[11]中利用假设检验方法，结合线性回归和聚类算法对股票价格曲线分段，每段时间区间对应于价格的上升期和下降期。接着将价格上升期与下降期内的相关新闻分别标注为有利新闻和不利新闻，用统计方法选择出新闻中的有利特征与不利特征。最后使用选择出的特征对股票价格做出预测。

TH Nguyen 等利用主题模型特征预测股票价格。他们在论文^[12]中提出了一种融合情感因素的主题模型，并在股票预测问题上取得了不错的效果。在金融新闻语料上，他们学习出情感主题模型，得到每篇新闻的主题分布向量。接着他们将新闻的主题分布向量作为特征，加入预测模型。由于模型在每个话题中都加入了情感维度参数，因此不同的主题能够反映新闻对于股价的利弊。最终他们的主题特征获得了不错的预测效果。

近几年来，学者们利用最新的计算机技术在股票预测领域取得了许多研究成果^[13-17]。尤其在自然语言处理领域，人们凭借深度学习技术取得了突破进展。丁效等将深度学习方法运用到股票预测领域。在论文^[18]中，他们提出了一种新的事件抽取方法，从新闻中抽取出结构化的事件。这些结构化的事件作为深度学习网络的输入，用于预测股票价格。随后，为了解决结构化事件中稀疏性的问题，在事件抽取工作的基础上，他们在论文^[19]中进一步学习出事件的嵌入式表达⁴。这些新闻事件的嵌入式表达被当作特征，作为卷积神经网络预测模型的输入。他们的股票预测方法取得了令人惊讶的效果。

2.2 情感分析

情感分析是自然语言处理的重要技术之一，它对带有情感色彩的主观文本进行分析处理，归纳出文本中蕴含的喜怒哀乐等情感倾向。情感分析方法是股票预测的一种手段，例如在论文^[10]中，作者通过分析大众媒体上群众的情感倾向预测股市走势，在论文^[11]中，作者将股票相关消息的情感元素作为预测股票涨跌的依据。这不单是因为股票相关新闻等市场信息中的观点会影响股价的波动，也是因为市场的散户在“从众心理”作用下往往跟从市场主流观点。

情感分析领域有着丰富的研究成果。早期的情感分析研究来源于对商品和电影等评论文本的观点分析。Turney 等人在论文^[20]中采用数学量互信息判断一个文本的情感极性。他们计算评论文本中一段短语与单词“excellent”和“poor”之间互信息的差值，以此探究商品评论的两极观点。Pang 等人在著作^[21]中实证化研究了情感分析在信息挖掘中的作用。在著作中，他们介绍了面向意见的信息搜索系统的实现方法，并重点探索了情感感知算法的解决方案。这些情感分析方法大多为基于统计学习的方法，通过计算数学统计量和假设检验方法以判断情感特征的作用。此外，为了方便未来情感分析方面的工作，他们还提供了相关的可用资源，基准数据集和评测活动。随着机器学习日渐成熟，许多学者将机器学习方法运用到情感分析领域。在论文^[22]中，Boiy 等人在博客、评论和论坛文本的情感分析中采用机器学习技术。他们手工标注部分文本，注释为正、负或中立的情感

⁴ 即 Embedding，对高维稀疏特征的低维化表示。

极性，并用于机器学习模型的训练。他们指出，在情感分析中引入机器学习方法的一大优势是跨语言模型的可移植性。2013 年，Mikolov 等人将单词的嵌入式表达（即 word embedding）引入自然语言处理领域^[23]。遗憾的是，Mikolov 等人提出的 word embedding 仅仅考虑了单词的语法特征，却难以区分单词的情感极性。随后，学者们开始研究针对情感分析问题的单词嵌入式表达。Tang 等人在面对推特文本情感分类问题时，提出了面向情感分析的 word embedding 学习方法^[24]。他们修改论文^[23]中的 CBOW 模型⁵，将语句的情感倾向和语法结构融合，最终学习出的 word embedding 既能反映单词的语法角色，又能体现单词的情感色彩。

2.3 长短时记忆网络（LSTM）

长短时记忆网络是一类具有特殊结构的循环神经网络⁶。它在网络结构中设计了具有存储状态的单元，因而适合处理时间序列中具有较长延迟与间隔的重要事件。

长短时记忆网络最早由 Hochreiter 等人在论文^[25]中提出。普通循环神经网络受到信号随时序指数衰减的影响，无法处理长时间间隔的信息关联问题。为了解决这一难题，Hochreiter 率先将“门”（Gate）与“记忆块”（Memory Block）的结构引入神经网络。LSTM 网络的每层都有多个相同结构的记忆块组成，每个记忆块，都包含一个“输入门”（Input Gate），一个“输出门”（Output Gate）和一个状态节点。输入门控制输入信息被记忆块接收的比例，输出门控制记忆块输出信息向外界传递的比例，而状态节点则蕴含记忆块所接收的所有历史信息，形成了对输入信息的“记忆”。凭借记忆块的结构，LSTM 网络能够记忆间隔超过 1000 个时间单位的信息。

Hochreiter 等人提出的 LSTM 结构，虽然能够记忆长时间间隔的信息，但是却存在“记忆爆炸”的问题。因为记忆块中的状态节点累积了所有的历史信息，所以随着时间的推移，状态节点的内部数值无限增长，状态节点因饱和而失效。Gers 等人在论文^[26]中通过引入“遗忘门”（Forget Gate）解决了节点饱和和失效的问题。在 Hochreiter 的 LSTM 结构中，状态节点此刻的信息完全传递到该节点下一刻的信息中，即状态节点不会“遗忘”它的历史信息。Gers 等人在状态节点上加入了“遗忘门”的结构，让节点此刻的历史信息按一定的比例传递到下一时刻。这一修改完善了 LSTM 的基本结构。现在人们使用的 LSTM 模型都包含了“遗忘门”的结构。

近几年来，学者们运用 LSTM 模型在自然语言处理领域取得了许多进展。

⁵ 即 continuous bag-of-words 模型，用于学习单词嵌入式表达的一种模型，另一种模型是 continuous skip-gram。

⁶ 循环神经网络，即 Recurrent Neural Network(RNN)。

Sundermeyer 等人使用 LSTM 单元构造语言模型，发现了它在语言模型上的潜力^[27]。在语素识别和混合语音识别问题上，Graves 等人设计出双向长短时记忆网络（Bidirectional LSTM），取得了最佳的识别效果^[28, 29]。

2.4 小结

本章介绍了论文的相关工作。2.1 节介绍了股票预测方面的研究，近几十年随着计算机技术的发展，学者们提出了许多不同的预测方法。2.2 节介绍了情感分析领域的发展，情感分析方法经常被运用于股票预测问题。2.3 节介绍了长短时记忆网络的研究与应用，本文提出的预测方法使用了该模型。

第三章 问题定义

股票预测问题是指利用股票在历史上的价格信息，相关的市场信息等，预测股票在未来一段时间内的价格走势或情况。根据价格信息的时间粒度，可分为以一天为时间粒度的低频交易，以及以小时、分钟甚至秒为时间粒度的高频交易。高频交易数据分析在工业界较为常见，其核心想法是通过分析大量的交易数据，学习股价涨跌前的交易信号规律^[30-33]。本文涉及的问题为低频交易问题，在学术界较为常见。其核心想法是通过辅之以相关新闻等市场信息，试图提高股票预测的准确性。

3.1 股票市场

本文研究的股票市场包括中国市场和美国市场。

中国市场主要研究“上证 A 股”市场，也即在上海证券交易所挂牌的一千余只股票。上证 A 股市场的每只股票均由国内上市公司（多为国有企业）发行，每个交易日⁷股票都可以自由交易。上海证券交易所的交易遵循 T+1 规则，当天买入的股票，当天不能卖出，需要等到下一个交易日才可以卖出这些股票。股票价格多数时间处于波动中，当股票的买入量多于卖出量，股票价格会上升，反之，当股票的卖出量多于买入量，股票价格会下降。在重大利好消息或利空消息刺激下，股价往往会暴涨或暴跌，加之股票之间相互关联，有时少数股票的暴涨暴跌会牵一发而动全身，因此股价的暴涨暴跌不利于股市的稳定。中国股市，例如上证 A 股市场，设置了 10% 的涨停与跌停限制，也即在股票价格超过上一个交易日价格的 10%，或者低于上一个交易日价格的 10% 后，停止该只股票当天的交易。

美国市场主要研究纳斯达克市场⁸（以下简称纳斯达克）与纽约证券交易所市场⁹（以下简称纽交所）上的股票。美国市场的股票多由世界闻名的大公司（例如阿里巴巴）发行，股票也可在每个交易日自由买卖。不同于国内市场，纳斯达克与纽交所的交易规则为 T+0，买入的股票随时可以卖出，无需等待。此外美国股票交易市场没有涨停与跌停限制，而是采用熔断机制¹⁰防止股价暴涨暴跌。

⁷ 除法定节假日外的周一至周五。

⁸ 即 Nasdaq(National Association of Securities Dealers Automated Quotations)，建立于 1971 年，是世界上第一个电子化证券市场。

⁹ 即 NYSE(New York Stock Exchange)，是上市公司总市值第一的交易所。

¹⁰ 在每季度开始，纽交所根据月前标普 500 指数平均值设定 7%、13% 和 20% 三个等级的熔断阈值，当 7% 和 13% 的跌幅发生时，会有 15 分钟的暂停交易时间，当跌幅触发 20% 时，交易将暂停直至休市。此外，个股也设有熔断机制，当某只股票交易价格在 5 分钟内涨跌幅超过 10% 时，暂停交易。

3.2 股票相关信息

市场中的相关信息，是推动股票价格波动的主要因素。例如，如果上市公司在财务报表中经营业绩良好的消息，会推动股价有所上涨。除了公司的业绩情况，上市公司的人事调整、重大决策等信息也会影响股价的涨跌。此外，国家财政政策与货币政策还会从宏观层面影响整个股票市场的走势。例如中央银行降低银行准备金率，这对于资本市场是一个利好消息，会促使资金流入股市，推动股价上涨。

实践中，我们从搜狐证券版块¹¹获取了与上证 A 股相关的新闻事件。我们首先抓取了搜狐证券版块的所有新闻，然后保留标题中包含股票名称的新闻，作为与该只股票相关的新闻。这些新闻中包含影响股价走势的信息，可以被利用于股票预测问题。对于纳斯达克与纽交所的相关新闻，我们从 SeekingAlpha 网站¹²上获取，该网站已经将相关新闻与股票对齐。该网站提供了美国股市股票的搜索功能，通过在网站搜索框输入股票代码，用户可以获取股票的每日价格、相关新闻和关键财务数据等信息。

3.3 股票涨跌预测问题

本文主要研究股票涨跌预测问题。股票在每个交易日结束后，都会留下一些关键交易数据：开盘价、收盘价¹³、最高价、最低价和交易量等。其中收盘价格一般用来代表这一交易日的股票价格。在持续的一段交易日内，股票会产生一个收盘价格的序列，我们记这段持续的价格序列为：

$$p_1, p_2, \dots, p_T$$

其中 p_t 表示股票在第 t 个交易日的收盘价格， T 为这段时间的交易日数量。

除了股票的价格信息，股票相关的市场信息也是股票预测的重要依据。假设在这段交易日内，该只股票相关的市场信息（相关新闻、消息等）为一组文本序列，每个交易日对应一个文本：

$$d_1, d_2, \dots, d_T$$

符号 d_t 表示股票在第 t 个交易日的相关市场信息。考虑到相关新闻可能会在非交易日出现，我们将非交易日的市场信息并入到最近的下一个交易日中。市场信息 d_t （相关新闻、消息等）是一组文档，可能包含多篇新闻报道。这组文档可以合并起来，看作单词的序列，形式化表示为：

¹¹ <http://stock.sohu.com/>

¹² <https://seekingalpha.com/>

¹³指股票在交易日里最后一笔买卖的成交价格。

$$d_t = w_{t1}w_{t2} \dots w_{tl_t}$$

其中 $w_{ti} \in V$ 是词汇表 V 中的一个单词, l_t 表示第 t 个交易日市场信息的长度。

股票价格相对前一个交易日会有所涨跌, 我们采用收盘价格作为涨跌的判断依据, 定义股票在第 t 个交易日的涨跌情况为 y_t , 若其值为 1 代表价格上涨, 若其值为 0 代表下跌或持平:

$$y_t = \begin{cases} 1, & \text{if } p_t > p_{t-1} \\ 0, & \text{if } p_t \leq p_{t-1} \end{cases}$$

于是股票涨跌预测问题定义如下:

已知一段持续的交易日内某只股票的价格序列和相关新闻序列:

$$\begin{aligned} p_1, p_2, \dots, p_T \\ d_1, d_2, \dots, d_T \end{aligned}$$

预测该只股票下一个交易日的涨跌情况, 即预测 y_{T+1} 。可以看出, 股票预测问题是利用股票历史信息预测未来涨跌的问题。

3.4 小结

本章介绍了本文研究的问题。3.1 节从买卖规则与交易限制等角度分别介绍了国内与国外的股票市场。在 3.2 节, 本文分析了一些能够有助于股票预测的相关信息。3.3 节以股票的价格信息与新闻信息为基础, 形式化定义了股票涨跌预测问题。

第四章 算法详述

4.1 算法综述

本文提出了一种面向股票预测的新闻特征抽取方法，并将 LSTM 模型运用于股票预测问题。算法的整体框架如图 4.1 所示。

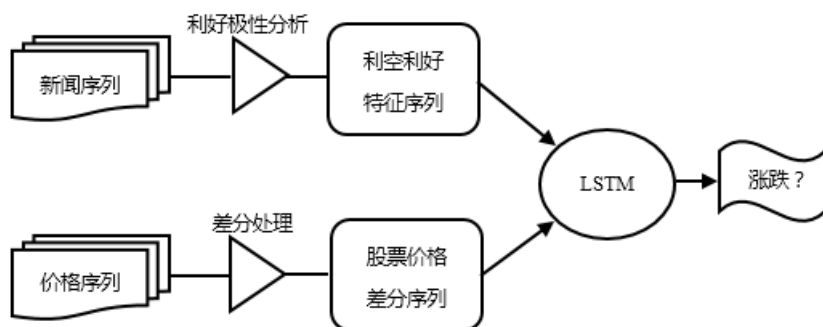


图 4.1 股票预测算法框架

模型的输入特征包括股票自身的价格特征与股票相关的新闻特征。价格特征方面，考虑到价格的绝对数据与涨跌绝对数据受到股价基数的影响，在不同股票上会有不同的数值大小，因此我们将股价进行差分处理，使用股价的涨跌幅作为价格特征。新闻特征抽取方面，我们提出了一种基于单词利好极性分析的特征抽取方法。首先，通过经验选取一组利好关键词与利空关键词。我们希望给每个单词定义一个“利好极性”¹⁴的数值，来判断一个单词在股票预测方面的情感极性。为此，我们采用最优化方法将利好关键词与利空关键词区分，使它们的利好极性数值之差尽可能大。通过解决最优化问题，可以获得词汇表中每个单词的利好极性，并以单词的利好极性为基础构造新闻的特征。新闻由单词序列构成，在词袋模型下，新闻可以看作单词在利好极性这一特征上的分布。我们将利好极性数值空间划分为若干个区间，通过统计方法获取新闻中单词在这些区间上的直方图分布，将直方图分布概率作为最终的新闻特征。

预测模型方面，我们选用多层 LSTM 模型。模型的输入是一段交易日内股票的价格特征与新闻特征，输出为下一个交易日的涨跌预测。选用 LSTM 模型有着直观的考虑：首先股票价格自身有着时序性的特征，每个交易日产生一期特征数据；其次，新闻事件等市场信息对股票价格的影响有延迟性与持续性，即新闻事件需要一定的时间间隔才能对市场产生影响，新闻事件对股价波动的影响往往会

¹⁴ 利好极性是一种非形式化的说法，类似于情感分析中单词的情感倾向。利好极性会在 4.2.2 节形式化定义。

持续一段时间。LSTM 模型作为一种循环神经网络模型，适合处理时序性问题，此外，由于 LSTM 自身的特殊结构，它相对其他神经网络模型更加适合处理带有延迟性与持续性特征的问题。

4.2 特征抽取

特征的构造与选择在许多机器学习问题中至关重要。股票涨跌预测问题是一个二分类问题，它的特征包括价格特征与新闻特征（市场信息特征）两个部分。

股票价格特征一般不直接使用收盘价格，因为不同股票的价格基数不同，收盘价格差异明显。我们对相邻交易日之间的收盘价格差分处理，获取收盘价格的涨跌比例，将它作为价格特征。新闻由文本构成，一般不直接作为模型输入。通过计算单词的利好极性数值，新闻文本形成了在利好极性数值上的单词分布。这一分布表征了新闻中的市场信息，能够作为新闻特征。

4.2.1 股票价格特征构造

股票价格特征是股票预测问题中最直接的、最直观的特征。相对于股价的绝对数值，股价的涨跌幅对于股价涨跌预测更为有效。不同的股票，由于股票价格基数不同，其股价的绝对值差异往往很大，但是股价的涨跌幅却很少受到价格基数的影响。上证 A 股市场设有涨停跌停限制，因此股价的涨跌幅在-10%到 10%之间。纳斯达克和纽交所虽然没有涨停跌停限制，但是股价的涨跌幅一般不会相差太大。

实验中，我们对股价差分处理，将每日股价相对于上一个交易日的涨跌幅作为当日的价格特征。

形式化的，股票收盘价格在一段持续的交易日内形成的价格序列为：

$$p_1, p_2, \dots, p_T$$

股价在第 i 个交易日相对上一个交易日的涨跌幅为：

$$q_i = \frac{p_i - p_{i-1}}{p_{i-1}}$$

差分处理后，股票的价格序列变为涨跌幅序列：

$$q_1, q_2, \dots, q_T \quad (4.1)$$

式子(4.1)中定义的股价涨跌幅序列是股票价格方面特征，作为后续 LSTM 预测模型的输入，如图 4.1 下半部分所示。

4.2.2 单词的利好极性

许多单词有着不同的情感倾向，例如“好”、“漂亮”属于褒义词，而“坏”、“丑陋”属于贬义词。在情感分析问题中，单词的情感倾向是重要的特征。比如，在推特文本情感分类问题中，一种简单的方法就是统计推文中褒义词与贬义词的数量，然后将两者的数量作为两个特征，训练一个诸如支持向量机（SVM）的二分类器。

不同于普通的情感分析，股票预测涉及到经济学的领域知识。一般而言，新闻对于股票价格的影响可分为利好和利空两类。利好新闻是指有利于股价上涨的消息。比如上市公司收购其他公司、公司项目中标、财报中净利润大幅增长等消息，这些消息反映出上市公司盈利能力改善，人们对公司股价的预期也会上升。相反，利空新闻是指导致股价下跌的消息。比如上市公司经营业绩不佳、内幕丑闻、与其他企业财务纠纷等消息，这些消息会导致人们抛售公司的股票。由此可见，股票预测中新闻利好与利空的判断，与情感分类问题类似，都有好坏两个方面，但是判断新闻的利好与利空由于涉及复杂的领域知识而更加困难。

本文提出一种能够结合领域知识的单词情感分析方法。假设我们有一个较大的与股票相关的新闻文档集合 D ，它包含 N 个新闻文档：

$$D = \{doc_1, doc_2, \dots, doc_N\} \quad (4.2)$$

这些新闻都与股票、上市公司相关，可以是网上对于上市公司的报道，也可以是金融机构对于某只股票的研究报告等。

每则新闻由单词序列构成，即文档切词¹⁵后的序列：

$$doc_i = w_1 w_2 \dots w_{n_i} \quad (4.3)$$

不同的单词在文档集中有不同的分布特征。有些单词会更多地在利好新闻中出现，有些单词会更多地在利空新闻中出现，而有些单词的分布则与新闻的种类没有关系。单词 w 在文档集 D 中出现的概率记为 $p(w)$ ，它表示包含单词 w 的文档在文档集合中的比例。符号 $pmi(w, v)$ 表示两个单词 w 和 v 之间的“点对互信息”^[34]，如式子(4.5)所示。

$$p(w) = \frac{|\{i | w \in doc_i\}|}{N} \quad (4.4)$$

$$pmi(w, v) = \ln \frac{p(w, v)}{p(w)p(v)} = \ln \frac{N * |\{i | w \in doc_i \text{ and } v \in doc_i\}|}{|\{i | w \in doc_i\}| * |\{i | v \in doc_i\}|} \quad (4.5)$$

点对互信息是数学统计量，表征两个单词的相关性。一般情况下，相关性大的单词，它们之间的点对互信息较大，相关性小的单词之间的点对互信息较小。

¹⁵ 切词，对于中文是 Segment，指的是将一个汉字序列切分成一个一个单独的词；对于英文是 Tokenize，指按照空格、标点等顺序地将英文文本切分为一个个单词。

例如单词“学校”与“老师”，它们很有可能在文档中共现，点对互信息较大，而单词“学校”与“工人”，它们的点对互信息则较小。

假设我们通过股票市场的经验，获取了一些关于股票的利好经验词与利空经验词。利好经验词是更有可能在利好新闻中出现的单词，例如“看好”，利空经验词是指那些更有可能在利空新闻中出现的单词，例如“亏损”。我们将这些单词的集合分别称为利好经验集 P_{exp} 和与利空经验集 N_{exp} ：

$$P_{exp} = \{u_1, u_2, \dots, u_p\} \quad (4.6)$$

$$N_{exp} = \{v_1, v_2, \dots, v_n\} \quad (4.7)$$

如果给单词赋予一个“利好极性”数值，使得在利好新闻中出现的单词极性值更大，那么利好经验集 P_{exp} 与利空经验集 N_{exp} 中单词的极性值之差应该尽可能大。一种可行的思路是试图找出一组参照单词，通过点对互信息定义单词表中单词的利好极性，使利好极性数值尽可能区分利好经验集 P_{exp} 与利空经验集 N_{exp} 中的单词。

以符号 P_{ref} 和符号 N_{ref} 分别表示利好参照集与利空参照集，如式子(4.8)和(4.9)所示。其中 P_{ref} 为一组数量为 K 的未知的利好参照单词， N_{ref} 为一组数量为 K 的未知的利空参照单词。式子(4.10)定义了单词在 P_{ref} 和 N_{ref} 上的利好极性，两组参照单词分别与利好新闻与利空新闻关联，通过点对互信息衡量了其他单词的利好极性。单词 w 的利好极性 $polar(w)$ 表征单词在利好新闻中出现的概率，与利好新闻的相关性；利好极性越大，单词在利好新闻中出现的概率越大，在利空新闻中出现的概率越小。

$$P_{ref} = \{w_{p1}, w_{p2}, \dots, w_{pK}\} \quad (4.8)$$

$$N_{ref} = \{w_{n1}, w_{n2}, \dots, w_{nK}\} \quad (4.9)$$

$$polar(w) \triangleq \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} pmi(w, v) - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} pmi(w, v) \quad (4.10)$$

为了找到一组最佳的利好参照集 P_{ref} 和利空参照集 N_{ref} ，将利好经验集 P_{exp} 与利空经验集 N_{exp} 区分开来，我们将上述的思路定义为最优化问题：

已知文档集合 D ，利好经验集 P_{exp} 和利空经验集 N_{exp} ，求解最优的参照集合 P^* 和 N^* ，其中 $polar(w)$ 由式子(4.10)定义， $p(w)$ 由式子(4.4)定义：

$$P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} \left[\frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} polar(w) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} polar(w) \right],$$

$$\text{subject to } |P_{ref}| = |N_{ref}| = K$$

$$\text{and } \forall w \in P_{ref}, p(w) \geq \varepsilon$$

$$\text{and } \forall w \in N_{ref}, p(w) \geq \varepsilon \quad (4.11)$$

参数 K 限定了参照集 P^* 和 N^* 的大小。参数 ε 是人为设置的频率阈值，防止在参照集中选入低频生僻词。

如何求解最优参照集 P^* 和 N^* ？我们先定义单词 w 在经验集 P_{exp} 和 N_{exp} 上的极性为：

$$polar_{exp}(w) \triangleq \frac{1}{|P_{exp}|} \sum_{v \in P_{exp}} pmi(w, v) - \frac{1}{|N_{exp}|} \sum_{v \in N_{exp}} pmi(w, v) \quad (4.12)$$

通过将式子(4.11)展开和变形，不难发现其中包含式子(4.12)定义的 $polar_{exp}(w)$ 。综合式子(4.10)(4.11)和(4.12)，我们得到（推导参见附录 D）：

$$\begin{aligned} P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} & \left[\frac{1}{K} \sum_{w \in P_{ref}} polar_{exp}(w) - \frac{1}{K} \sum_{w \in N_{ref}} polar_{exp}(w) \right], \\ \text{subject to } & |P_{ref}| = |N_{ref}| = K \\ \text{and } & \forall w \in P_{ref}, p(w) \geq \varepsilon \\ \text{and } & \forall w \in N_{ref}, p(w) \geq \varepsilon \end{aligned} \quad (4.13)$$

利好经验集 P_{exp} 和利空经验集 N_{exp} 中的单词是已知的，因此对于单词表 V 中的任意单词 w ，式子(4.12)中定义的 $polar_{exp}(w)$ 是确定的。同时，由于文档集 D 已知，因此单词 w 出现的概率 $p(w)$ 也是确定的。如果我们过滤去单词表中在文档集 D 中出现概率小于 ε 的单词，将剩余的 M 个单词按照 $polar_{exp}(w)$ 的数值从大到小排序，即：

$$polar_{exp}(w_{e1}) > polar_{exp}(w_{e2}) > polar_{exp}(w_{e3}) > \dots > polar_{exp}(w_{eM}) \quad (4.14)$$

那么从式子(4.13)中容易发现，最优化参照集 P^* 为序列(4.14)中的前 K 个单词，而最优化参照集 N^* 为序列(4.14)中的后 K 个单词。至此，我们找到了最优化问题(4.11)的最优解：

$$\begin{aligned} P^* &= \{w_{e1}, w_{e2}, \dots, w_{eK}\} \\ N^* &= \{w_{e(M-K+1)}, w_{e(M-K+2)}, \dots, w_{eM}\} \end{aligned} \quad (4.15)$$

最优参照集 P^* 和 N^* 能够将经验集 P_{exp} 和 N_{exp} 区别开来，同样也能界定其他单词的利好极性。最优参照集本身也是一种参照集，我们最终在 P^* 和 N^* 上定义单词的利好极性，作为后续新闻特征抽取的基础，如式子(4.16)所示。

$$polar(w) \triangleq \frac{1}{K} \sum_{v \in P^*} pmi(w, v) - \frac{1}{K} \sum_{v \in N^*} pmi(w, v) \quad (4.16)$$

纵观单词的利好极性分析，该方法既能赋予单词一个关于利好与利空的评分，又能结合股票市场相关的领域知识。该方法有一个需要人工控制的关键点，即利好经验集 P_{exp} 和利空经验集 N_{exp} 的构造。这需要结合经济学知识与股票市场经验，通过浏览大量的相关新闻中来确定。在后续的实验部分，我们会给出单词利好极性分析的直观效果。

4.2.3 股票新闻特征构造

文本向量化有多种方式。最简单的方法是词袋模型（Bag of words），它将文档映射为单词表长度的向量。向量的每个维度对应单词表中的一个单词，每个维度的大小对应该单词在文档的频数。词袋模型可以计算文档之间的相似度，但是存在向量维度大，内容稀疏的缺点。主题模型是文档向量化的另一种方式。在主题模型中，文档被看作在若干个隐含主题上的概率分布，而每个主题又是词汇表上单词的概率分布。Blei 等提出的 LDA 主题模型^[35]，是最经典的主题模型之一。LDA 主题模型是一种无监督学习方法，能将文档降维到主题向量上。这种文档表示方法虽然保留了文档的主题与内容信息，却因为忽视了情感倾向而在股票预测问题中效果一般。

考虑到股票预测这一特殊问题，我们将 4.2.2 节中单词的利好极性融入到新闻特征构造中。在式子(4.16)中，我们在最优参照集 P^* 和 N^* 上定义了单词的利好极性。如果我们以单词的利好极性数值为变量，就能统计文档中单词在利好极性上的频率分布。这一频率分布就是文档在利好利空方面的一种特征。由于频率分布的维度过高（维度大小为单词表长度），我们将利好极性按数值划分为若干区间，用直方图频率分布去近似最初的频率分布，既降低文档特征的维度，又保留了分布的真实性。这一方法的形式化描述如下：

假设在单词表 V 中，单词利好极性的最大值为 Max (4.17)，最小值为 Min (4.18)。将利好极性数值范围平均划分为 L 个区间，每段对应一个利好极性区间 I (4.19)，与直方统计图的区间一一对应：

$$Max = \max_{w \in V} polar(w) \quad (4.17)$$

$$Min = \min_{w \in V} polar(w) \quad (4.18)$$

$$I_j = \left[Min + (j - 1) * \frac{Max - Min}{L}, Min + j * \frac{Max - Min}{L} \right], j \in \{1, 2, \dots, L\} \quad (4.19)$$

对于式子(4.3)中的第 i 个交易日的新闻序列 $doc_i = w_1 w_2 \dots w_{n_i}$ ，它的特征向量 $f(doc_i)$ 定义为：

$$f(doc_i) = (x_{i1}, x_{i2}, \dots, x_{iL})^T$$

$$\text{其中 } x_{ij} = \frac{1}{n_i} |\{k | \text{polar}(w_k) \in I_j\}| \quad (4.20)$$

可见，式子(4.20)中定义的特征向量的每一维对应直方图的一个区间，其值表示文档中单词在该利好极性区间的频率。参数 L 既是直方图区间的数量，又是新闻特征向量的维数。 L 的数值需要在实验中调整，因为如果 L 太小，会让特征向量失去信息，而如果 L 太大，特征向量将失去降维的效果，向量每个维度的方差会很大。

式子(4.20)所示的新闻特征向量一方面是对新闻文本的抽象，起到了低维表示的效果，另一方面，特征向量结合了单词的利好极性，直观反映出新闻对于股票价格的利好利空影响。

对于股票涨跌预测问题中的相关新闻序列 d_1, d_2, \dots, d_T ，经过单词利好极性分析与新闻特征抽取，我们获得了这些新闻的特征向量序列：

$$f(d_1), f(d_2), \dots, f(d_T) \quad (4.21)$$

我们将这一特征序列作为后续 LSTM 预测模型的输入，如图 4.1 上半部分所示。

4.3 预测模型

在股票涨跌预测问题中，不论是股票每日的价格信息，还是相关的市场信息，都具有时间维度上的顺序。股票每个交易日会产生收盘价，股票的相关新闻也能对应到某个交易日。相对于其他的分类算法，例如逻辑斯蒂回归^[36]与支持向量机^[37]，循环神经网络模型能够自然地融入时间这一维度，描述特征在时序上的关联。

除了时序性，股票预测问题还有延迟性与持续性特征。首先从相关新闻的发布到新闻影响股价的波动有一定的时间间隔。在新闻发布后，人们从发现新闻，传播新闻，到分析新闻，做出决策需要时间，新闻信息作用到股价上需要一段时间。其次新闻对股价的影响往往不局限于单个交易日，而是在一段交易日内产生持续的作用。股票预测延迟性与持续性的特征，要求预测模型能够将新信息保存一段时间，具有一定的“记忆”能力。长短时记忆网络（LSTM），如同它的名称一样，能够将输入信息记忆在模型中，是处理股票预测问题的合适模型。

4.3.1 LSTM 层

记忆块（Memory Block）是长短时记忆网络的基本构成单元。记忆块如同大脑中的神经元，具有记忆信息与连接周围神经元的作用。记忆块的结构如图 4.2 所示，

记忆块与时序相关，在 t 时刻它的输入是 x_t ，输出是 h_t ，记忆块“记忆”的历史信息为 C_t 。每个记忆块都有三个门，即输入门 i_t ，遗忘门 f_t 以及输出门 o_t 。输入门控制 t 时刻输入信息的比例，遗忘门控制 t 时刻记忆历史信息的比例，输出门则决定了记忆块向下一层输出信息的比例。

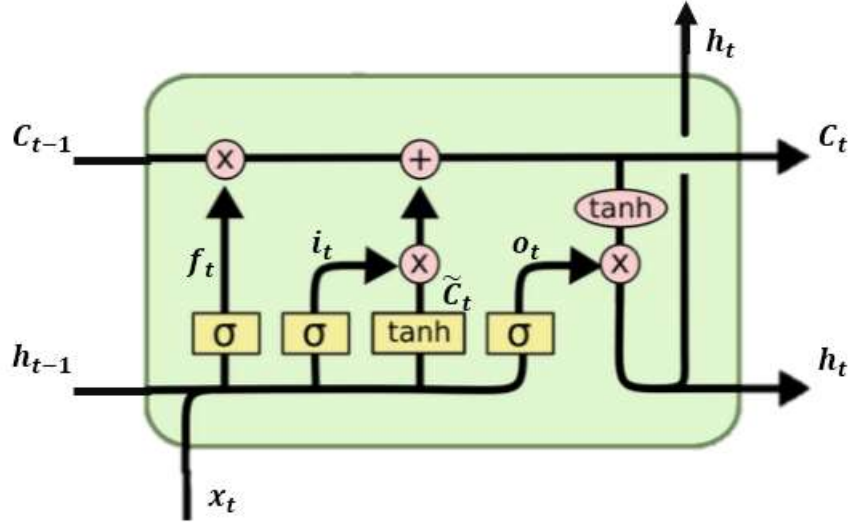


图 4.2 Memory Block of LSTM

记忆块具有时序的状态， C_t 包含了从时刻 0 到时刻 t 的所有输入信息，是记忆块在时刻 t 的状态。

遗忘门 f_t 由输入信息 x_t 和记忆块上一时刻的输出信息 h_{t-1} 决定，经过 sigmoid 函数¹⁶变换，遗忘门的数值在 0 到 1 之间，如式子(4.22)所示。 f_t 表示记忆块在前一时刻的状态 C_{t-1} 有多大比例保留到当前 t 时刻。 f_t 值为 0 表示完全遗忘前一时刻的状态， f_t 值为 1 表示完全记忆前一时刻的状态。

输入信息 x_t 与前一时刻的输出 h_{t-1} ，经过 tanh 函数¹⁷变换，形成当前时刻的状态增量 \tilde{C}_t ，如式子(4.23)所示。状态增量 \tilde{C}_t 的值在区间 $(-1,1)$ ，表示输入信息 x_t 能够给记忆块的状态造成的增量大小。

与遗忘门类似，输入门 i_t 的值由输入信息 x_t 和前一时刻的输出信息 h_{t-1} 决定，如式子(4.24)所示。输入门控制着状态增量 \tilde{C}_t 被记忆块接收的比例。如果输入门 i_t 值为 0，那么状态增量 \tilde{C}_t 会被完全忽略；如果输入门 i_t 值为 1，那么 \tilde{C}_t 会被完全计入状态 C_t 。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4.22)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4.23)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4.24)$$

¹⁶ sigmoid 函数将实数区间 \mathbb{R} 映射到区间 $(0,1)$ ，一般用符号 σ 表示， $\sigma(x) = 1/(1 + e^{-x})$ 。

¹⁷ tanh 函数即双曲正切函数，将实数区间 \mathbb{R} 映射到区间 $(-1,1)$ ， $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$

现在我们来查看记忆块的状态是如何更新的。状态值 C_t 由两部分组成，一个部分是前一时刻的状态 C_{t-1} ，这部分由遗忘门控制保留的比例，另一个部分是状态增量 \tilde{C}_t ，这一部分由输入门决定接收增量的比例。状态值 \tilde{C}_t 随着时间的更新公式如(4.25)所示。值得注意的是，公式(4.25)中的符号“*”表示向量之间的按位乘法¹⁸，即在记忆块中门是对信息的一种缩放。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4.25)$$

输出门 o_t 与遗忘门和输入门类似，其值由输入信息 x_t 与前一时刻的输出信息 h_{t-1} 综合而来，如式子(4.26)所示。输出门 o_t 的大小决定了记忆块的状态 C_t 有多大概率作为输出，被其他的神经网络层捕获。

表 4.1 LSTM 层伪代码

LSTM 层: $lstm_layer(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T)$
输入: 一组按时序排列的向量组 $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T$;
参数张量 $\vec{\varphi} = [W_i, W_f, W_o, W_c, \vec{b}_i, \vec{b}_f, \vec{b}_o, \vec{b}_c]$
输出: 按时序排列的向量组 $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_T$
过程:
1. $\vec{C}_0 := \vec{0}$
2. $\vec{h}_0 := \vec{0}$
3. for t from 1 to T:
4. $\vec{i}_t := \sigma(W_i \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_i)$
5. $\vec{f}_t := \sigma(W_f \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_f)$
6. $\vec{o}_t := \sigma(W_o \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_o)$
7. $\vec{C}_t := \tanh(W_c \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_c)$
8. $\vec{C}_t = \vec{f}_t * \vec{C}_{t-1} + \vec{i}_t * \vec{C}_t$
9. $\vec{h}_t = \vec{o}_t * \tanh(\vec{C}_t)$
10. endfor
11. return $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_T$

符号 h_t 表示了记忆块在时刻 t 的输出。这一输出基于记忆块的状态 C_t ，但是需要经过输出门的过滤。状态 C_t 首先经过 \tanh 函数变换，将状态压缩到区间 $(-1,1)$ ，在由输出门 o_t 决定输出的比例，如式子(4.27)所示。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4.26)$$

$$h_t = o_t * \tanh(C_t) \quad (4.27)$$

¹⁸ 按位乘法即向量每个维度相乘，最后得到的是同维的向量，例如 $(1,2,3)*(4,5,6)=(4,10,18)$ 。

LSTM 记忆块是一种时序上的函数变换，它将输入时序序列变换为另一个时序序列。因为记忆块内置状态单元，并且具有输入门与遗忘门的特殊结构，所以记忆块具备关联长间隔信息的能力。

注意到，记忆块的输入 x_t 和状态 C_t 都是向量，也就是说，记忆块可以单独作为神经网络的一层。LSTM 网络由一层或多层构成，我们将一层 LSTM 网络的功能用伪代码表示出来，如表 4.1 所示。在后续的章节中，我们用符号“lstm_layer()”表示一层 LSTM 网络。

从表 4.1 可以看出，LSTM 层的输入与输出都是时序上的向量组，并且输入向量与输出向量在时间上一一对应。每个时刻，LSTM 层都进行一组相同的操作，计算出记忆块的状态和输出信息。多层 LSTM 模型由多个 LSTM 层拼接而成。

4.3.2 LSTM 预测模型

从图 4.1 中可以看出，预测模型需要利用价格信息与新闻信息两种不同的特征。

价格特征如式子(4.1)所示，由每个交易日收盘价的涨跌幅组成。价格信息具有较强的时效性，股价情况与近几个交易日的涨跌较为相关，与相隔较远的交易日关联微弱。因此在模型中，我们使用的价格特征为近 3 个交易日的涨跌幅。

新闻特征序列如式子(4.21)所示。新闻特征具有延迟性与持续性特征，它对股价的影响往往会持续一段交易日。本文采用双层 LSTM 模型来处理股票的新闻特征。结合价格特征与新闻特征的预测模型如图 4.3 所示。

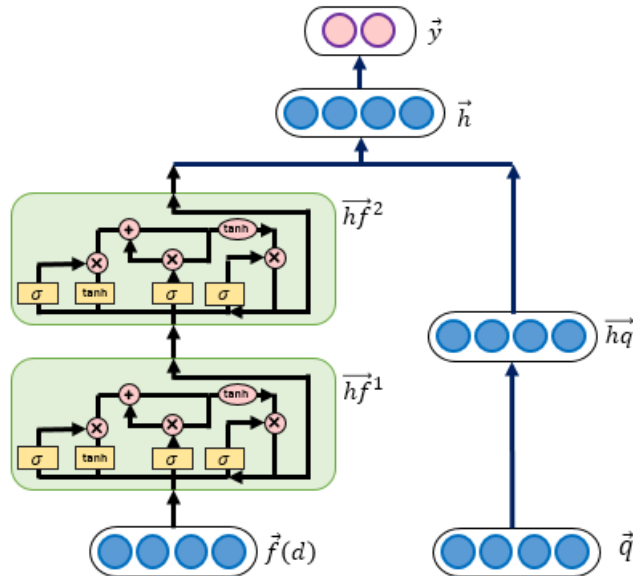


图 4.3 LSTM 预测模型

预测模型由两部分组成，模型的左半部分处理新闻特征序列，主要由两个

LSTM 层构成。模型的右半部分是股票的价格特征，经过一层神经网络的处理。新闻信息与价格信息在更高的网络层汇合，共同用于预测股价的涨跌。

LSTM 模型预测过程的伪代码如表 4.2 所示。输入包括输入特征与模型参数两个部分。输入特征包含股票前 a 个交易日的涨跌幅序列 \vec{q} （实验中，我们设置 a 的值为 3），即利用股价前 a 个交易日的涨跌幅特征。输入特征还包括股票前 b 个交易日的新闻特征序列 \vec{f} 。该序列由 b 个特征向量构成，每个特征向量都由式子(4.20)定义，并且按顺序对应这 b 个交易日。模型参数源自神经网络的每一层，这些参数通过训练过程获得（在 4.3.3 节中介绍），在预测过程中保持不变。

预测过程的输出是下一个交易日股价的涨跌概率。符号 \hat{y}_{T+1} 代表第 $T+1$ 个交易日股价的涨跌概率，它具有两个维度，分别对应下跌与上涨的概率。本文用符号 $\hat{x}[i]$ 表示向量 \hat{x} 的第 i 个维度。因此 $\hat{y}_{T+1}[0]$ 与 $\hat{y}_{T+1}[1]$ 分别表示股价下跌与上涨的概率。

表 4.2 LSTM 模型预测过程伪代码

LSTM 模型的预测过程
<p>输入： 式子(4.1)中前 a 个交易日的股价涨跌幅序列 $\vec{q} = (q_{T-a+1}, q_{T-a+2}, \dots, q_T)^T$，式子 (4.21) 中前 $b(b>a)$ 个交易日新闻的特征向量序列 $\vec{f}(d_{T-b+1}), \vec{f}(d_{T-b+2}), \dots, \vec{f}(d_T)$；模型参数矩阵 W_q, W_h, W_y，偏移向量 $\vec{b}_q, \vec{b}_h, \vec{b}_y$，两个 LSTM 层的参数张量 $\vec{\varphi}^1, \vec{\varphi}^2$</p> <p>输出： 第 $T+1$ 个交易日下跌和上涨的概率 \hat{y}_{T+1}</p> <p>过程：</p> <ol style="list-style-type: none"> $\vec{hf}_T^1, \vec{hf}_{T-b+1}^1, \dots, \vec{hf}_T^1 := lstm_layer(\vec{f}(d_{T-b+1}), \vec{f}(d_{T-b+2}), \dots, \vec{f}(d_T))$ $\vec{hf}_T^2, \vec{hf}_{T-b+1}^2, \dots, \vec{hf}_T^2 := lstm_layer(\vec{hf}_T^1, \vec{hf}_{T-b+1}^1, \dots, \vec{hf}_T^1)$ $\vec{hq} := ReLU(W_q \cdot \vec{q} + \vec{b}_q)$ $\vec{h} := ReLU(W_h \cdot [\vec{hq}, \vec{hf}_T^2] + \vec{b}_h)$ $\vec{y} := Softmax(W_y \cdot \vec{h} + \vec{b}_y)$ return ($\vec{y}[0], \vec{y}[1]$)

模型的预测过程与图 4.3 中的流程一致。第一步，新闻的特征向量序列 \vec{f} 经过一个 LSTM 层变化为向量序列 \vec{hf}^1 ，LSTM 层的变换过程如表 4.1 所示。第二步，向量序列 \vec{hf}^1 再经过一层 LSTM 转化为向量序列 \vec{hf}^2 。多层 LSTM 变换的意义在于捕获新闻特征序列中高维的特征，这也是深度学习区别于普通机器学习的一大特点。第三步，价格特征序列 \vec{q} 经过一个全连接层变化为隐含向量 \vec{hq} ，该向量包含了价格特征的信息。第四步，第二步中向量序列 \vec{hf}^2 中最后一个向量 \vec{hf}_T^2 ，与第

三步中的隐含向量 \vec{h}_q ，组合成一个新的向量，经过一个全连接层变化为隐含向量 \vec{h} 。向量 $\vec{h}f_T^2$ 累积了这 b 个交易日的所有新闻特征，代表近 b 个交易日市场方面的综合信息。隐含向量 \vec{h}_q 则对应了股票近 a 个交易日价格信息。向量 \vec{h} 综合 $\vec{h}f_T^2$ 与 \vec{h}_q 的内容，蕴含着股票的价格信息与市场信息。第五步，向量 \vec{h} 经过线性变换与 Softmax 函数¹⁹变换，得到输出层 \vec{y} 。输出层仅包含两个节点，分别对应股票价格下跌和上涨的概率。第六步，返回涨跌预测的结果，输出层向量 \vec{y} 是二维向量， $\vec{y}[0]$ 表示预测下跌的概率， $\vec{y}[1]$ 表示预测上涨的概率。

神经网络的全连接层都包含两个变换，输入向量经过线性变换和非线性变换得到输出向量。线性变换一般为矩阵变换，而非线性变换（激活函数）却有多种，比如 sigmoid 函数与 tanh 函数。在模型中，本文使用的变换为 ReLU 函数²⁰，因为 ReLU 函数能够有效解决深度神经网络中梯度消失的问题^[38]。

4.3.3 模型参数学习

LSTM 预测模型中的参数需要经过训练得到。如表 4.2 的输入部分所示，模型的参数包括参数矩阵 W_q, W_h, W_y ，偏移向量 b_q, b_h, b_y ，以及 LSTM 层的参数张量 $\vec{\varphi}^1, \vec{\varphi}^2$ 。我们将这些参数集合为整个预测模型的参数，用符号 θ 表示：

$$\theta = \{W_q, W_h, W_y, \vec{b}_q, \vec{b}_h, \vec{b}_y, \vec{\varphi}^1, \vec{\varphi}^2\} \quad (4.28)$$

为了学习参数 θ ，需要定义模型的损失函数。股票涨跌预测问题是二分类问题，适宜使用交叉熵损失函数。假设在一段长为 $T+1$ 的交易日内，观测到某只股票的价格序列为 $\vec{p} = p_1, p_2, \dots, p_{T+1}$ ，相关新闻序列为 $\vec{d} = d_1, d_2, \dots, d_{T+1}$ 。经过特征抽取，由式子(4.1)和式子(4.21)获得价格涨跌序列 q_1, q_2, \dots, q_{T+1} 和新闻特征向量序列 $f(d_1), f(d_2), \dots, f(d_{T+1})$ 。如式子(4.29)所示，由涨跌序列 q_1, q_2, \dots, q_{T+1} 容易得到价格涨跌的标签向量 $\vec{y}_1, \vec{y}_2, \dots, \vec{y}_{T+1}$ 。假设预测模型（如表 4.2 所示）对第 $b+1$ 个交易日至第 $T+1$ 个交易日的涨跌预测结果为 $\hat{y}_{b+1}, \hat{y}_{b+2}, \dots, \hat{y}_{T+1}$ 。

$$\vec{y}_i = \begin{cases} (1, 0) & \text{if } q_i \leq 0 \\ (0, 1) & \text{if } q_i > 0 \end{cases} \quad i = 1, 2, \dots, T+1 \quad (4.29)$$

模型的交叉熵损失函数定义为：

$$L_c(\vec{p}, \vec{d}; \theta) = -\sum_{i=b+1}^{T+1} \sum_{k=0}^1 \vec{y}_i[k] \ln \hat{y}_i[k] \quad (4.30)$$

过拟合（overfitting）是机器学习中常见的问题。模型的损失函数可以分解为

¹⁹ Softmax 函数是神经网络输出层的常用变换，一般用于多分类问题。它将 K 维向量 x 变换为 K 维向量 $\text{Softmax}(x)$ ，其中 $\text{Softmax}(x)_j = e^{x_j} / \sum_i e^{x_i}$ ， $j = 1, 2, \dots, K$

²⁰ ReLU，即 Rectified Linear Units， $\text{ReLU}(x) = \max(0, x)$

偏差与方差两个部分^[39]。当模型参数太多，结构太复杂时，损失函数中方差部分会变大，导致模型虽然在训练集上拟合效果很好，但是在测试集上的泛化错误率却很高。这种现象被称为过拟合。解决过拟合现象的常用方法是在损失函数中加入参数的正则化项。正则化方法适用于大多数神经网络模型，例如全连接网络、卷积神经网络等，但是不适合 LSTM 层网络。LSTM 的记忆块需要保留输入历史信息，而在梯度下降算法中，正则化是一种忽视部分输入信息的手段，这两者南辕北辙。“Dropout”方法由 Srivastava 等人在论文^[40]中提出，是另一种解决神经网络过拟合的常用方法，在 LSTM 网络中较为常见。“Dropout”方法通过让神经网络中的节点按一定的概率失效（失效即该节点不接受前一层节点的信息，不向后一层节点传递信息），防止神经网络过拟合。

如图 4.3 所示的 LSTM 预测模型包括 LSTM 层与全连接层两种类型的网络层。我们采用“Dropout”与参数正则化相结合的方法防止模型过拟合。在 LSTM 层 \vec{hf}^1 与 \vec{hf}^2 ，采用“Dropout”方法，以一定的概率 α 让记忆块中的节点失效。其他全连接层的参数则被加入到损失函数中。最终，LSTM 预测模型的损失函数如式子 (4.31) 所示。

$$Loss(\vec{p}, \vec{d}; \theta) = - \sum_{i=b+1}^{T+1} \sum_{k=0}^1 \tilde{y}_i[k] \ln \hat{y}_i[k] + \beta \|\theta'\|^2$$

$$with \theta' = \{W_q, W_h, W_y, \vec{b}_q, \vec{b}_h, \vec{b}_y\} \quad (4.31)$$

式子 (4.31) 所示的损失函数包含两个部分，前一部分是交叉熵损失函数 $L_c(\vec{p}, \vec{d}; \theta)$ ，表示模型预测错误的代价，后一部分是全连接层参数的正则化项，采用岭回归（Ridge Regression）方法，代表对模型复杂度的惩罚。参数 β 是模型的超参数，表征正则化项在损失函数中的比重。

模型的参数学习采用 RMSProp^[41] 算法。RMSProp 算法是对随机梯度下降 (SGD)^[42] 算法的一种改进，它能缓冲 SGD 算法中梯度的剧烈变化。LSTM 预测模型的代码实现使用了 Keras²¹ 库。

4.4 小结

本章详细阐述了本文提出的股票预测方法。4.1 节概述了算法的流程。4.2 节介绍了股票预测特征抽取方法，价格信息方面，本文以价格涨跌幅度取代价格绝对值作为特征，市场信息方面，本文提出了一种基于单词利好极性的新闻特征抽取方法。4.3 节首先介绍了 LSTM 网络层，然后提出了一种结合新闻特征的 LSTM 预测模型，最后给出了该模型的损失函数和学习方法。

²¹ <https://keras.io/>

第五章 实验

5.1 数据集

实验数据从性质上分为价格数据和新闻数据两类，从市场角度分为中国市场数据与美国市场数据两种。

5.1.1 数据采集与处理

数据采集即数据抓取的过程。互联网上有丰富的股票相关数据。想要获取它们，首先要人工搜索包含数据的站点，然后利用网页爬虫抓取这些网页，最后分析网页结构，挖掘出股票相关数据。

本文研究的中国市场指上证 A 股市场，研究的美国市场指纳斯达克与纽交所。股票相关数据来源如表 5.1 所示。上证 A 股市场股票的每日价格信息可以从中国财经站点²²抓取。上证 A 股市场的相关新闻数据来源于搜狐财经版块²³与投资之家²⁴。纳斯达克与纽交所股票的价格信息与相关消息都获取自 SeekingAlpha 网站。

表 5.1 股票相关数据来源

数据类别	来源网站	来源网址例子
上证 A 股价格数据	http://finance.china.com.cn/stock/	http://app.finance.china.com.cn/stock/quote/history.php?code=sh601766&type=daily
上证 A 股新闻数据	http://stock.sohu.com/	http://q.stock.sohu.com/news/cn/766/601766/4932863.shtml
纳斯达克与纽交所价格数据	https://seekingalpha.com/	https://seekingalpha.com/symbol/BABA/key-data/historical-quotes
纳斯达克与纽交所新闻数据	https://seekingalpha.com/	https://seekingalpha.com/symbol/BABA/news

利用网页爬虫与网页分析工具，可以从表 5.1 所示的网站获得股票相关数据。这些数据已共享到 GitHub 上，数据的情况在附录 C 中有详细介绍。我们以获取纽交所股票阿里巴巴（股票代码 BABA）价格数据和新闻数据为例，介绍数据采集的过程。

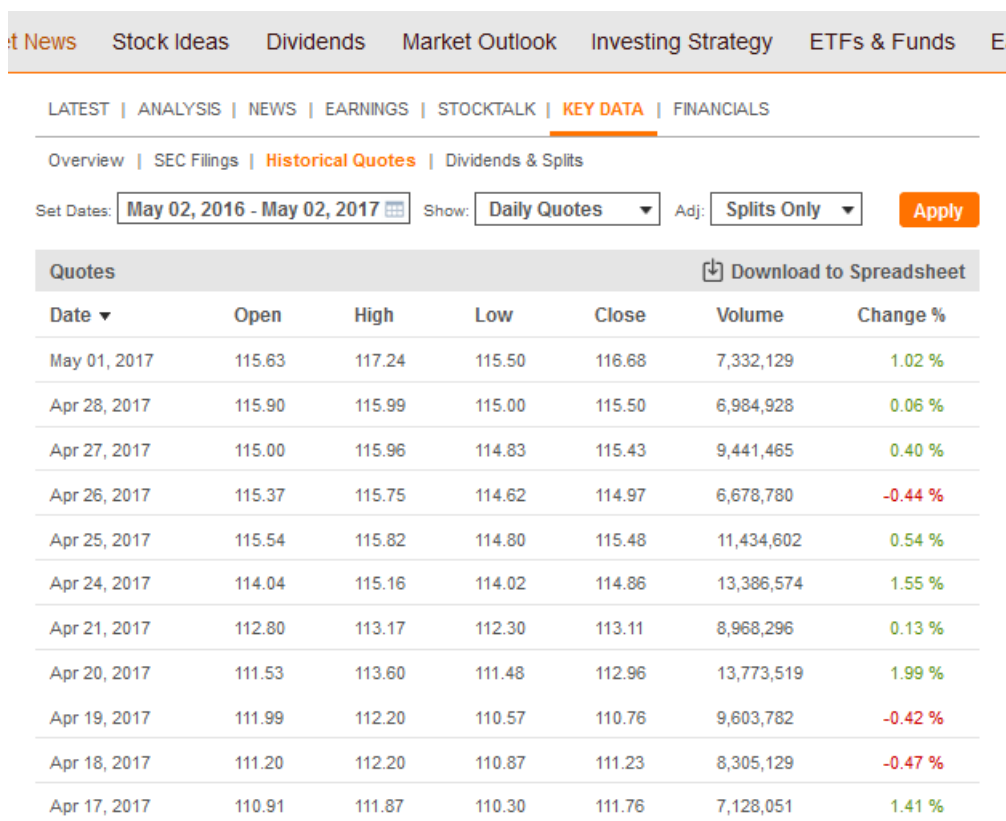
SeekingAlpha 网站通过在网页地址（URL）中加入股票代码来区分不同的股

²² <http://finance.china.com.cn/stock/>

²³ <http://stock.sohu.com/>

²⁴ <http://finance.china.com.cn/>

票。股票的价格信息在网页子目录“KEY DATA”下，阿里巴巴价格信息的网页地址为“<https://seekingalpha.com/symbol/BABA/key-data/historical-quotes>”，网页主体部分如图 5.1 所示。网页上半部分是股票的价格走势图，没有在图 5.1 中显示，网页下半部分是股票交易信息表，如图 5.1 的主体部分。表的每行对应一个交易日，每列对应一项交易信息。交易信息包括开盘价(Open)、最高价(High)、最低价(Low)、收盘价(Close)、换手量(Volume)和当日涨跌幅(Change)。



Quotes Download to Spreadsheet						
Date ▼	Open	High	Low	Close	Volume	Change %
May 01, 2017	115.63	117.24	115.50	116.68	7,332,129	1.02 %
Apr 28, 2017	115.90	115.99	115.00	115.50	6,984,928	0.06 %
Apr 27, 2017	115.00	115.96	114.83	115.43	9,441,465	0.40 %
Apr 26, 2017	115.37	115.75	114.62	114.97	6,678,780	-0.44 %
Apr 25, 2017	115.54	115.82	114.80	115.48	11,434,602	0.54 %
Apr 24, 2017	114.04	115.16	114.02	114.86	13,386,574	1.55 %
Apr 21, 2017	112.80	113.17	112.30	113.11	8,968,296	0.13 %
Apr 20, 2017	111.53	113.60	111.48	112.96	13,773,519	1.99 %
Apr 19, 2017	111.99	112.20	110.57	110.76	9,603,782	-0.42 %
Apr 18, 2017	111.20	112.20	110.87	111.23	8,305,129	-0.47 %
Apr 17, 2017	110.91	111.87	110.30	111.76	7,128,051	1.41 %

图 5.1 SeekingAlpha 网站中股票的数据页面

这张交易信息表是关键，主要抓取步骤如下：

第一步，连接目标网页。程序使用 Selenium²⁵工具包，启动一个模拟浏览器，程序模拟浏览器操作，打开阿里巴巴股票的网页，并在输入框中填入需要查询的交易日期范围。这时，浏览器会向 SeekingAlpha 网站发出数据请求，爬虫线程短暂睡眠后，浏览器与 SeekingAlpha 网站通信完成，获得了需要的数据。浏览器页面此时会显示这段交易日期范围内股票的每日交易信息。

第二步，获取页面代码。模拟浏览器提供了获取页面代码的接口，程序从浏览器获得目标网页的 HTML 代码，这段代码以 HTML 格式存储了网页的页面内容，其中包含交易信息表。

²⁵ Selenium 是一个自动化浏览器工具，可用于爬取网页数据，网址 <http://www.seleniumhq.org/>

第三步，解析目标网页。程序调用 Jsoup²⁶工具包，解析 html 代码的结构。程序利用 HTML 文本中的 tag, id 和 class 等属性，在 HTML 的 DOM Tree²⁷中定位交易信息表，从表格中获得股票的交易信息，最后将表格中的信息记录到本地文件。

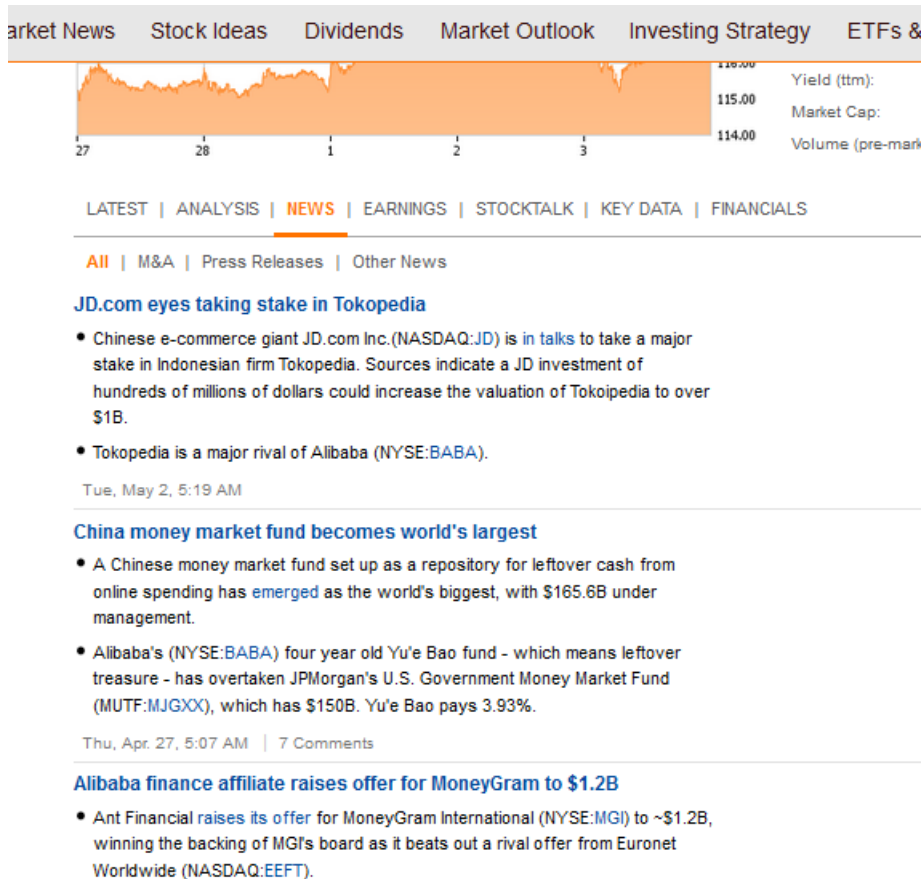


图 5.2 SeekingAlpha 网站中股票的相关新闻页面

阿里巴巴股票的相关新闻在网页子目录“NEWS”下，其相关新闻网页地址为“https://seekingalpha.com/symbol/BABA/news”，网页主体部分如图 5.2 所示。网页上半部分依然是股票的价格走势图，在图 5.2 中显示了一小部分，网页下半部分是股票相关消息的列表，如图 5.2 的主体部分。这些消息按时间顺序由近及远排列，每则消息的格式都相同，消息与消息之间通过分栏符相隔。一则消息由标题，内容和时间等元素构成。标题以蓝色大号字体显示在最前面，言简意赅，往往由十个左右的单词组成。消息的内容位于消息的中间，以黑色小号字体显示，相对于标题更加详实，长度为几十个单词左右。时间位于消息的最下方，以灰色小号字体显示，表明了这则消息发布的具体时间。

图 5.2 所示的相关新闻在 HTML 代码中是一张表格，也可以使用 Selenium 与

²⁶ Jsoup 是一个 Java 的 HTML 解析器，网址 <https://jsoup.org/>

²⁷ DOM 是 Document Object Model 的简称。DOM Tree 将 HTML 页面按标签解析成树形结构。

Jsoup 工具包获得。与获取价格数据的步骤相似，程序首先根据 URL 地址连接目标网页，然后获取网页的 HTML 代码，最后利用 Jsoup 工具包解析页面代码，得到表格中的所有相关消息。前两步的具体实现两者基本一致，但是在最后一步中，前者比后者要更为复杂。

```

    ▶ <li id="anew_3260371" class="mc_list_li"></li>
    ▶ <li class="mc_line_li mc_new" style="display:none"></li>
    ▼ <li id="anew_3257201" class="mc_list_li">
        ▶ <div class="ticker_date_left"></div>
        ▼ <div class="mc_list_texting right bullets">
            ▼ <div class="mc_bullets_title mc_summaries_title_link">
                ▼ <a class="market_current_title" href="/news/3257201-alibaba-finan
                    target="_self"> ev
                    Alibaba finance affiliate raises offer for MoneyGram to $1.2B
                </a>
            </div>
            ▶ <span class="general_summary light_text bullets"></span>
            ▼ <span class="mc_new">
                <span class="date pad_on_summaries">Mon, Apr. 17, 3:27 AM</span>
            </span>
            <span class="mc_gray_separator">|</span>
            <span class="market_current_comment">4 Comments</span>
        </div>
        <div class="cleaner"></div>
    </li>
    ▶ <li class="mc_line_li mc_new" style="display:none"></li>
    ▶ <li id="anew_3256622" class="mc_list_li"></li>
    
```

图 5.3 新闻页面部分 HTML 代码

图 5.3 显示了图 5.2 中第三则消息“Alibaba finance affiliate raises offer for MoneyGram to \$1.2B”页面部分近邻的 HTML 代码。图中左侧的每个灰色三角形代表网页 DOM Tree 中的一个节点，对应 HTML 代码中的一个标签。朝右的三角表示未展开子节点的节点，朝下的三角表示已展开子节点的节点。从代码中可以看到，每则消息都是表格的一行，即一个“li”标签。第三则消息的标签“li”的 id 属性是“anew_3257201”，class 属性是“mc_list_li”。消息的标题是一个超链接标签（标签“a”），标签的 class 属性是“market_current_title”。消息的内容对应一个区域标签（标签“span”），标签的 class 属性为“general_summary”、“light_text”和“bullets”。消息的时间也对应一个区域标签，区域内的文本为时间信息，该标签的 class 属性为“date”和“pad_on_summaries”。页面解析程序利用 HTML 代码中的标签类别（如上文提到的表格标签“li”，超链接标签“a”和区域标签“span”等），以及这些标签的属性特征（例如 id 属性和 class 属性）来定位和获取消息的标题、内容和时间等重要信息。

Selenium 工具能够模拟网页的打开、点击、输入与选择等操作，便于爬虫程

序应对复杂的数据显示流程。Jsoup 工具支持 HTML 文本的高效解析，对于从页面代码中寻找关键数据很有帮助。基于这两个工具包，数据采集程序能从互联网获得各类股票相关数据，包括中国股市与美国股市，数据的与文本的，如表 5.1 所示。

从搜狐证券版块抓取的新闻多种多样，只有小部分与上证 A 股市场的股票相关。对于搜狐新闻，我们保留标题中含有股票名称的新闻，将新闻与股票对齐，而过滤去其他的无关新闻。对于美国股市的新闻，因为 SeekingAlpha 网站已经将新闻与相应股票关联，所以无需额外对齐处理。

5.1.2 数据集统计

为了计算单词的利好极性（4.2.2 节），我们选取了较大的文档集进行计算。对于中文单词，文档集由 52922 篇股票相关新闻组成，对于英文单词，文档集由 24969 篇股票相关新闻构成。

实验重点关注那些相关新闻较多的股票。表 5.2 和表 5.3 统计了实验中使用的股票数据的交易天数（正常交易的交易日数量）和相关新闻天数（有新的相关新闻的交易日数量）。由于股票数量较多，我们没有一一列出每只股票的统计信息。表 5.2 包含了 12 只上证 A 股股票的平均信息，表 5.3 包含了 34 只著名的纳斯达克与纽交所股票的平均信息。中国股市股票数据的时间范围为 2013 年 1 月 1 日至 2015 年 10 月 14 日，其间包含 671 个交易日。美国股市数据集包括了从 2012 年 9 月 1 日到 2016 年 10 月 31 日的股票信息，其间包含 1047 个交易日。

表 5.2 实验使用的中国股票相关信息统计

股票数量	时间跨度	平均交易日数	平均相关新闻天数
12	2013.1.1 至 2015.10.14	622.08	191.67

表 5.3 实验使用的美国股票相关信息统计

股票数量	时间跨度	平均交易日数	平均相关新闻天数
34	2012.9.1 至 2016.10.31	1022.85	411.32

5.2 实验结果

5.2.1 单词利好极性

单词的利好极性分析是新闻特征抽取的基础。表 5.4 展示了极性值处于最两

端的单词。从表中可以看出，极性值最大的几个单词与“增长”与“收益”相关，这些单词往往在利好消息中出现；而极性值最小的几个单词与“违法”相关，这些单词往往在利空消息中出现。

表 5.4 极性值处于两端的单词

单词集	内容
极性值最大的 10 个中文单词	{优势、有望、EPS、提升、加快、需求、维持、创新、增速、互联网}
极性值最小的 10 个中文单词	{违规、违反、涉嫌、证监局、责令、立案、华锐、处罚、决定书、违法}
极性值最大的 10 个英文单词	{dividend, properties, yield, realty, growth, trust, payable, income, REIT, in-line}
极性值最小的 10 个英文单词	{trial, lawsuit, damages, file, judge, rule, request, court, sue, ruling}

5.2.2 股票预测

股票涨跌预测实验中，本文将每只股票的数据（价格信息与相关新闻信息）按照时间顺序排列，将前 80% 的数据作为训练集，后 20% 的数据作为测试集。以测试集上涨跌预测的准确率作为模型优劣的标准。

本文设计了几种不同的预测方法，与论文中的模型进行比较。第一种方法将前 3 个交易日价格的涨跌幅作为特征，以 SVM 模型为预测模型，这种方法记为 SVM-Price。第二种方法的特征是上一个交易日价格的涨跌幅，使用的模型是循环神经网络，这种方法记为 RNN-Price。第三种方法将前 3 个交易日价格的涨跌幅以及相关新闻作为特征，使用 SVM 模型做出预测，这种方法记为 SVM-News。第四种方法为本文第 4 章提出的预测方法，记为 LSTM-News。四种方法在 12 只国内股票和 34 只美国股票上的预测准确率如表 5.5 和表 5.6 所示。

表 5.5 国内股价涨跌预测准确率比较

股票代码	SVM-Price	RNN-Price	SVM-News	LSTM-News
600028	0.5581	0.5581	0.5814	0.5930
600030	0.5075	0.5150	0.5448	0.5522
600315	0.5512	0.5275	0.6063	0.5970
600519	0.5448	0.5522	0.5672	0.5746
600887	0.5373	0.5149	0.5228	0.5672
601318	0.4851	0.5298	0.5224	0.5672

601628	0.4801	0.4925	0.4925	0.5659
601857	0.5176	0.5176	0.5176	0.5224
601998	0.5149	0.5149	0.4925	0.5000
600597	0.5167	0.5083	0.5083	0.5299
600138	0.5895	0.5448	0.6045	0.5113
600383	0.5223	0.5373	0.5074	0.4925
平均	0.5271	0.5261	0.5390	0.5478

从表 5.5 中可以看出, 在 12 只上证 A 股股票上, 方法一(SVM-Price)与方法二(RNN-Price)的效果相近, 分别有 52.71%与 52.61%的准确率。方法三(SVM-News)在 12 只股票上的平均准确率为 53.90%, 相对于方法一有 1.19%的提升。方法四(LSTM-News)的准确率为 54.78%, 相对于方法一提升了 2.07%, 在四种方法中效果最好。

表 5.6 美股股价涨跌预测准确率比较

股票代码	SVM-Price	RNN-Price	SVM-News	LSTM-News
AAL	0.5068	0.4932	0.5274	0.5342
AIG	0.4976	0.5167	0.5263	0.5215
AMZN	0.5502	0.5455	0.5646	0.5598
AXP	0.5263	0.5263	0.5120	0.5407
BABA	0.5047	0.4860	0.4953	0.5140
BA	0.5455	0.5550	0.5789	0.5885
BAC	0.4928	0.5502	0.5263	0.5455
BHI	0.5694	0.5263	0.5359	0.5550
BIDU	0.4928	0.4833	0.5072	0.4785
BRK.A	0.5455	0.5263	0.5742	0.6268
BRK.B	0.5502	0.5359	0.5694	0.6411
COF	0.4833	0.5072	0.5502	0.5502
COV	0.5502	0.5502	0.5981	0.5742
DUK	0.5598	0.5598	0.5789	0.5407
EBAY	0.5598	0.5598	0.5311	0.5455
FDX	0.5072	0.5215	0.5502	0.5455
GE	0.4833	0.4880	0.5407	0.5742
GM	0.5072	0.5072	0.5455	0.5502
GOOG	0.5167	0.4833	0.5407	0.5455
HD	0.4976	0.4976	0.5072	0.5072
HPQ	0.5311	0.5024	0.5598	0.5789
JNJ	0.5072	0.5072	0.5407	0.5598
JPM	0.4737	0.5024	0.5311	0.5407

KO	0.5167	0.4976	0.5311	0.5359
MOT	0.5120	0.5263	0.5455	0.5550
MS	0.5263	0.5167	0.5502	0.5455
ORCL	0.5024	0.5024	0.5359	0.5407
PEP	0.5311	0.5359	0.5359	0.5789
PG	0.5167	0.4928	0.5694	0.5598
SINA	0.5024	0.5072	0.5120	0.4928
T	0.5024	0.5072	0.5407	0.5502
WMT	0.5311	0.5407	0.5646	0.5550
XOM	0.5263	0.5263	0.4976	0.5407
YHOO	0.5359	0.5359	0.5263	0.5550
平均	0.5195	0.5182	0.5412	0.5508

在 34 只美股股票上，方法一(SVM-Price)与方法二(RNN-Price)的预测效果也很相近，分别有 51.95%与 51.82%的准确率。方法三(SVM-News)的准确率为 54.12%，相对于方法一与方法二有所提高。方法四(LSTM-News)在四种方法中效果最佳，准确率为 55.08%，相对于方法一有 3.14%的提升。

综合表 5.5 与表 5.6 中的结果，可以发现本文提出的新闻特征有助于预测股票涨跌，这种特征在结合 LSTM 预测模型后效果更好。但是，从实验中也可以看出，这种新闻特征在少部分股票上没有效果甚至效果适得其反（例如 BIDU，SINA）。本文认为，股票预测是一项复杂的任务，受到社会中随机事件的影响，预测算法很难做到适用于所有的股票。

5.2.3 投资回报率

模拟交易实验是另一种检验股票预测的方法。模拟交易的一种评价标准为日均回报率^[43]，如式子(5.1)所示。

$$r = \frac{1}{T} \sum_{t=1}^T \frac{y_t - y_{t-1}}{y_{t-1}} \quad (5.1)$$

实验中，模拟交易采用的股票买卖策略参考了论文^[44]中的方法，由(a)和(b)两条策略构成：

(a)如果预测下个交易日股价会涨，那么今天收盘时买入。如果下个交易日能够盈利 1.5%，那么立即卖出，否则股票在收盘时抛出。

(b)如果预测下个交易日股价会跌，那么今天收盘时做空卖出。如果下个交易日能够盈利 1.5%，那么立即买入，否则收盘时买入股票。

本文在 34 只美国股票上做了模拟交易实验，实验结果如表 5.7 所示。

表 5.7 34 只美股上模拟交易实验结果

方法	SVM-Price	RNN-Price	SVM-News	LSTM-News
平均日均回报率	0.005791%	-0.00169%	0.01769%	0.02214932%

从实验结果看，方法四的收益率最高，方法三与方法四的收益率要高于方法一和方法二。从相对数据来看，本文提出的新闻特征对于股票预测有一定的效果。

值得注意的是，四种方法的日均收益率均偏低，最高的只有 0.0221%，方法二的日均收益率甚至为负数。考虑到实盘操作中股票买卖需要交易费用，因此算法距离实际应用还有一定距离。

5.3 小结

本章 5.1 节介绍了数据采集与处理过程，并对数据集进行统计。作者使用爬虫工具与页面解析工具，从互联网上抓取了中国股市与美国股市的相关数据。5.2 节给出了算法在数据集上的效果。在股票涨跌预测实验中，本文提出的方法与对照算法相比有 2%到 3%的准确率提升，在股票模拟交易实验中，本文的方法有助于提高回报率，但是与实际应用还有一定距离。

第六章 结论与未来工作

6.1 结论

本文在股票预测方面主要有两个贡献。

一是提出了一种基于单词利好极性分析的新闻特征抽取方法。单词的利好极性分析方法结合了股票市场的经验,能够区分出代表利好新闻与利空新闻的单词。新闻特征是新闻中单词的利好极性值的分布,反应了新闻对于股票价格的影响。

二是设计了一种用于股票预测的长短时记忆网络模型。该模型将新闻特征与价格特征结合,考虑了股票价格的时序性与相关新闻影响的持续性。

实验中,在 12 只上证 A 股股票上,相对基于价格特征的 SVM 方法,本文提出的预测方法能将预测准确率提高 2.08%;在 34 只纳斯达克与纽交所股票上,本文提出的方法给准确率带来 3.14%的提升。在 34 只美国股票上的模拟交易实验显示,本文提出的新闻特征有助于提高投资回报率,但是距离实际运用还有一定差距。

6.2 未来工作

本文提出的新闻特征是通过人工方式提取的,虽然有一定的效果,但是在抽取过程中丢弃了诸多的信息。深度学习具有特征学习的能力,是替代手工特征提取的一种方式。作者曾尝试基于单词的 word2vec,使用卷积神经网络处理相关新闻,并结合循环神经网络构造预测模型。该模型首先使用卷积神经网络从相关新闻中提取出最显著的文本特征,然后将连续交易日的文本特征作为循环神经网络的输入。遗憾的是模型效果不佳。作者觉得问题的可能原因在于 word2vec 忽视了单词的情感倾向,因此准备在后续工作中,结合股票市场学习单词的嵌入式表达,并运用于股票预测。此外,本文提出的新闻特征抽取方法基于词袋模型,粒度较粗。新闻中最具有价值的内容是“事件”,“事件”是人们对股价走势进行判断的重要依据。因此,后续工作拟运用自然语言处理的技术从新闻中提取“事件”级别的信息,实现信息的精准定位,作为特征抽取的基础。

参考文献

- [1] Fama E F. The behavior of stock-market prices[J]. The journal of Business, 1965, 38(1): 34-105.
- [2] Li Q, Jiang L L, Li P, et al. Tensor-Based Learning for Predicting Stock Movements[C]//AAAI. 2015: 1784-1790.
- [3] Schumaker R P, Chen H. A quantitative stock prediction system based on financial news[J]. Information Processing & Management, 2009, 45(5): 571-583.
- [4] Fung G P C, Yu J X, Lam W. News sensitive stock trend prediction[M]//Advances in knowledge discovery and data mining. Springer Berlin Heidelberg, 2002: 481-493.
- [5] Markowitz H. Portfolio selection[J]. The journal of finance, 1952, 7(1): 77-91. MLA
- [6] Chan S W K, Franklin J. A text-based decision support system for financial sequence prediction[J]. Decision Support Systems, 2011, 52(1): 189-198.
- [7] Schumaker R P, Zhang Y, Huang C N, et al. Evaluating sentiment in financial news articles[J]. Decision Support Systems, 2012, 53(3): 458-464.
- [8] Kim K, Han I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index[J]. Expert systems with Applications, 2000, 19(2): 125-132.
- [9] Sehgal V, Song C. Sops: stock prediction using web sentiment[C]//Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. IEEE, 2007: 21-26.
- [10] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.
- [11] Fung G P C, Yu J X, Lam W. Stock prediction: Integrating text mining approach using real-time news[C]//Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on. IEEE, 2003: 395-402.
- [12] Nguyen T H, Shirai K. Topic modeling based sentiment analysis on social media for stock market prediction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015.
- [13] Akita R, Yoshihara A, Matsubara T, et al. Deep learning for stock prediction using numerical and textual information[C]//Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, 2016: 1-6.
- [14] Chen W, Hao Z, Cai R, et al. Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction[J]. Soft Computing, 2016, 20(11): 4575-4588.
- [15] Day M Y, Lee C C. Deep learning for financial sentiment analysis on finance news providers[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016: 1127-1134.

- [16] Gao T, Li X, Chai Y, et al. Deep learning with stock indicators and two-dimensional principal component analysis for closing price prediction system[C]//Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on. IEEE, 2016: 166-169.
- [17] Kim K, Yang S, Kim D, et al. A Stock Prediction System Based on News and Twitter[J]. International Journal of Software Engineering and Its Applications, 2016, 10(6): 69-80.
- [18] Ding X, Zhang Y, Liu T, et al. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation[C]//EMNLP. 2014: 1415-1425.
- [19] Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (ICJAI). 2015: 2327-2333.
- [20] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.
- [21] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends? in Information Retrieval, 2008, 2(1-2): 1-135.
- [22] Boiy E, Moens M F. A machine learning approach to sentiment analysis in multilingual Web texts[J]. Information retrieval, 2009, 12(5): 526-558.
- [23] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [24] Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]//ACL (1). 2014: 1555-1565.
- [25] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [26] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. Neural computation, 2000, 12(10): 2451-2471.
- [27] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[C]//Interspeech. 2012: 194-197.
- [28] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.
- [29] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 273-278.
- [30] Kirilenko A A, Kyle A S, Samadi M, et al. The flash crash: High frequency trading in an electronic market[J]. 2016.
- [31] Menkveld A J. High frequency trading and the new market makers[J]. Journal of Financial Markets, 2013, 16(4): 712-740.
- [32] Nayak R, Te Braak P. Temporal pattern matching for the prediction of stock prices[C]//Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining-Volume 84. Australian Computer Society, Inc., 2007: 95-103.

- [33] Brogaard J, Hendershott T, Riordan R. High-frequency trading and price discovery[J]. *Review of Financial Studies*, 2014, 27(8): 2267-2306.
- [34] Bouma G. Normalized (pointwise) mutual information in collocation extraction[J]. *Proceedings of GSCL*, 2009: 31-40.
- [35] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. *Journal of machine Learning research*, 2003, 3(Jan): 993-1022.
- [36] Menard S. *Applied logistic regression analysis*[M]. Sage, 2002.
- [37] Cortes C, Vapnik V. Support vector machine[J]. *Machine learning*, 1995, 20(3): 273-297.
- [38] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//*Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010: 807-814.
- [39] Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma[J]. *Neural computation*, 1992, 4(1): 1-58.
- [40] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. *Journal of Machine Learning Research*, 2014, 15(1): 1929-1958.
- [41] Tieleman T, Hinton G. Lecture 6.5-RMSProp, COURSE: Neural networks for machine learning[J]. University of Toronto, Tech. Rep, 2012.
- [42] Bottou L. Large-scale machine learning with stochastic gradient descent[M]//*Proceedings of COMPSTAT'2010*. Physica-Verlag HD, 2010: 177-186
- [43] Hellstrom T, Holmstrom K. Predicting the stock market[J]. Unpublished Thesis, Malardalen University, Department of Mathematics and Physics, Vasteras, Sweden, 1998.
- [44] Fung G P C, Yu J X, Lam W. News sensitive stock trend prediction[C]//*Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2002: 481-493.

附录 A 符号表

下表列出了本文使用的符号及其含义。

符号	意义
p_t	股票第 t 个交易日的收盘价
d_t	股票第 t 个交易日的相关新闻文档
w	单词表 V 中的单词
V	单词表
y_t	股票第 t 个交易日的涨跌标签，1 表示上涨，0 表示下跌或持平
q_t	股票第 t 个交易日的涨跌幅
D	新闻文档集合
$pmi(w, v)$	单词 w 和 v 的点对互信息
P_{exp}, N_{exp}	利好经验集与利空经验集
P_{ref}, N_{ref}	利好参照集与利空参照集
P^*, N^*	最优利好参照集与利空参照集
$polar(w)$	单词 w 的利好极性
$f(d_t)$	第 t 个交易日新闻的特征向量
$lstm_layer()$	一层 LSTM 网络
\hat{y}_t	模型预测的第 t 个交易日股价的涨跌概率向量
\vec{y}_t	股票第 t 个交易日的涨跌标签向量
W	模型的参数矩阵
\vec{b}	模型的参数偏移量
$\vec{h}, \vec{hf}, \vec{hq}$	模型的中间层
θ	预测模型的参数集合
r	模拟交易实验中的日均回报率

附录 B 硕士期间科研成果

会议论文

股票预测：一种基于新闻特征抽取和循环神经网络的方法

会议名称：全国信息检索学术会议(CCIR)，2016

本人排名：1

导师排名：3

附录 C 股票数据集

硕士期间，我从互联网上抓取了许多股票相关的数据集。有的是中国股市的，有的是美国股市的；有的是交易信息，有的是相关新闻与政策。下表列出了这些数据的文件名称，内容简介和共享链接。

名称	简介	链接
<i>CnNewsReport</i>	上证 A 股所有股票 3 年的相关新闻和研究报告	https://github.com/zeyazhang/sewm_stock_data/blob/master/CnNewsReport.rar
<i>CnStockPrice</i>	上证 A 股所有股票 3 年的日交易信息	https://github.com/zeyazhang/sewm_stock_data/blob/master/CnStockPrice.rar
<i>PeopleDaily</i>	《人民日报》要闻版块 7 年的新闻内容	https://github.com/zeyazhang/sewm_stock_data/blob/master/PeopleDaily.rar
<i>UsStockNews</i>	172 只美国股票 4 年的相关新闻	https://github.com/zeyazhang/sewm_stock_data/blob/master/UsStockNews.rar
<i>UsStockPrice</i>	172 只美国股票 4 年的日交易信息	https://github.com/zeyazhang/sewm_stock_data/blob/master/UsStockPrice.rar
<i>MainUsStockPrice</i>	649 只纳斯达克股票 10 年的日交易信息	https://github.com/zeyazhang/sewm_stock_data/blob/master/MainUsStockPrice.rar
<i>UsMarketNews</i>	SeekingAlpha 网站 4 年的股票相关消息	https://github.com/zeyazhang/sewm_stock_data/blob/master/SeekingAlphaUsStockNews.rar

附录 D 部分公式推导

从式子 (4.11) 到式子 (4.13) 的推导如下:

已知:

$$polar(w) \triangleq \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} pmi(w, v) - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} pmi(w, v) \quad (4.10)$$

$$\begin{aligned} P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} & \left[\frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} polar(w) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} polar(w) \right], \\ \text{subject to } & |P_{ref}| = |N_{ref}| = K \\ \text{and } \forall w \in P_{ref}, & p(w) \geq \varepsilon \\ \text{and } \forall w \in N_{ref}, & p(w) \geq \varepsilon \end{aligned} \quad (4.11)$$

$$polar_{exp}(w) \triangleq \frac{1}{|P_{exp}|} \sum_{v \in P_{exp}} pmi(w, v) - \frac{1}{|N_{exp}|} \sum_{v \in N_{exp}} pmi(w, v) \quad (4.12)$$

求证式子 (4.11) 等价于式子 (4.13), 即

$$\begin{aligned} P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} & \left[\frac{1}{K} \sum_{w \in P_{ref}} polar_{exp}(w) - \frac{1}{K} \sum_{w \in N_{ref}} polar_{exp}(w) \right], \\ \text{subject to } & |P_{ref}| = |N_{ref}| = K \\ \text{and } \forall w \in P_{ref}, & p(w) \geq \varepsilon \\ \text{and } \forall w \in N_{ref}, & p(w) \geq \varepsilon \end{aligned} \quad (4.13)$$

证明:

首先,

$$pmi(w, v) = \ln \frac{p(w, v)}{p(w)p(v)} = \ln \frac{p(v, w)}{p(v)p(w)} = pmi(v, w)$$

式子 (4.11) 中需要最大化的部分变形如下:

$$\begin{aligned} & \frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} polar(w) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} polar(w) \\ &= \frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} \left[\frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} pmi(w, v) - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} pmi(w, v) \right] \end{aligned}$$

$$\begin{aligned}
 & -\frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} \left[\frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} pmi(w, v) - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} pmi(w, v) \right] \\
 & = \frac{1}{|P_{exp}| |P_{ref}|} \sum_{w \in P_{exp}} \sum_{v \in P_{ref}} pmi(w, v) - \frac{1}{|P_{exp}| |N_{ref}|} \sum_{w \in P_{exp}} \sum_{v \in N_{ref}} pmi(w, v) \\
 & - \frac{1}{|N_{exp}| |P_{ref}|} \sum_{w \in N_{exp}} \sum_{v \in P_{ref}} pmi(w, v) + \frac{1}{|N_{exp}| |N_{ref}|} \sum_{w \in N_{exp}} \sum_{v \in N_{ref}} pmi(w, v) \\
 & = \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} \frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} pmi(w, v) - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} \frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} pmi(w, v) \\
 & - \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} pmi(w, v) + \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} pmi(w, v) \\
 & = \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} \left[\frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} pmi(w, v) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} pmi(w, v) \right] \\
 & - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} \left[\frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} pmi(w, v) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} pmi(w, v) \right] \\
 & = \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} [polar_{exp}(v)] - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} [polar_{exp}(v)]
 \end{aligned}$$

将式子 (4.11) 中的限制条件 $|P_{ref}| = |N_{ref}| = K$ 代入, 将符号 v 替换为 w 可得:

$$\begin{aligned}
 & \frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} polar(w) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} polar(w) \\
 & = \frac{1}{|P_{ref}|} \sum_{v \in P_{ref}} [polar_{exp}(v)] - \frac{1}{|N_{ref}|} \sum_{v \in N_{ref}} [polar_{exp}(v)] \\
 & = \frac{1}{K} \sum_{w \in P_{ref}} polar_{exp}(w) - \frac{1}{K} \sum_{w \in N_{ref}} polar_{exp}(w)
 \end{aligned}$$

此式即为式子 (4.13) 中需要最大化的部分。

因此，式子 (4.11) 中的问题

$$\begin{aligned}
 P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} & \left[\frac{1}{|P_{exp}|} \sum_{w \in P_{exp}} \text{polar}(w) - \frac{1}{|N_{exp}|} \sum_{w \in N_{exp}} \text{polar}(w) \right], \\
 \text{subject to } & |P_{ref}| = |N_{ref}| = K \\
 \text{and } \forall w \in P_{ref}, & p(w) \geq \varepsilon \\
 \text{and } \forall w \in N_{ref}, & p(w) \geq \varepsilon
 \end{aligned} \tag{4.11}$$

等价于式子 (4.13) 中的问题：

$$\begin{aligned}
 P^*, N^* = \operatorname{argmax}_{P_{ref}, N_{ref}} & \left[\frac{1}{K} \sum_{w \in P_{ref}} \text{polar}_{exp}(w) - \frac{1}{K} \sum_{w \in N_{ref}} \text{polar}_{exp}(w) \right], \\
 \text{subject to } & |P_{ref}| = |N_{ref}| = K \\
 \text{and } \forall w \in P_{ref}, & p(w) \geq \varepsilon \\
 \text{and } \forall w \in N_{ref}, & p(w) \geq \varepsilon
 \end{aligned} \tag{4.13}$$

证毕.

致谢

2015年3月，我在朋友的劝说下买入了中国北车的股票，2个月内，股价经历了暴涨暴跌，曾经一度翻倍，最后又回落到起点。后来才发现，那段时间正是中国股市动荡，机遇与风险并存的时期。在实际操盘中，我感觉到关注股市是一件费时的事情，萌生了能否利用计算机实盘操作的想法。在导师闫宏飞老师的鼓励下，我开始了对股票市场的研究。

从研一暑假开始，我从互联网上不停地抓取股票相关数据，上证市场的，纳斯达克市场的，价格数字的，新闻文本的等等。数据抓取是一个繁琐的任务，为了获取不同的数据源需要编写不同的爬虫程序。在这一过程中，我渐渐成为一个熟练的爬虫使用者。股票预测问题涉及机器学习与自然语言处理技术，我在探索的过程中，得到了来自实验室老师和同学的大力帮助。闫宏飞老师对股票预测研究很感兴趣，时常与我讨论相关研究，并提出一些建议。王锦鹏师兄帮助我理解了 word2vec 相关技术，陈维政师兄对主题模型颇有研究，不仅向我推荐相关论文，还经常提供可能的研究方向。天网组的读书会活动也拓宽了我的技术视野，夯实了我的理论基础。我尤其感谢李睢，他对于机器学习理论的理解很深，经常将深奥的理论层层剖析，加深了我对于机器学习技术的理解。在大家的帮助和自己的努力下，研二的暑假我在 CCIR 上发表了股票预测的论文，虽然只是国内会议，但是也是对我这段时间努力付出的肯定。

在探索股票预测方法的过程中，我觉得收获最大的并不是发表论文，完成毕业设计，而是自己在自然语言处理，机器学习技术方面的积累。因为在以后的工作中，我并不会去专门研究股票预测，但是所用到的计算机技术却能成为我的财富。

我的北大时光即将走到尽头，马上就要踏入工作岗位。我收获很多，也留下了许多遗憾。北大“兼容并包，思想自由”的办学方针，让我在北大认识了许多不拘一格的同学，北大学识渊博的老师和文理兼备的课程，帮助我拓宽了视野，丰富了学识。回顾在北大的七年，我觉得最大的遗憾是没有交到几个像兄弟一样的朋友，没有在校园找到生命的另一半。时光从身旁穿过，成熟多了，当年那个大清早爬起来背诵四级单词，让室友一脸惊愕的学生已不复存在，当年那个写了一沓情诗，偷偷放到女神座位的青涩少年已不复存在，留下的只是偶尔回想起自己会忍俊不禁的记忆。北大是一个很高的平台，给了我许多的机会，但是当近在咫尺的机会从指间滑落，我也曾感到无比的挫败。本科的时期放弃出国留学，研究生的时候部委面试失败，是我觉得人生轨迹受到最大影响的两次经历。然而自哀人生无益，我只能总结经验，收拾起心情匆匆上路，未来还有许多抉择需要我去

面对。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校 ☐ 一年 / ☐ 两年 / ☐ 三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日