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 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。传统股票价格的预测大多局限于银行板块数据系列，股票收盘价，月度加权股值等 [1]。

但是最近的研究表明，公共领域的大量在线信息，如主流媒体的新闻，社交媒体讨论，政府发布的官方文案都可能会对投资者对金融市场的看法产生很大的影响 [3]。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。目前已经有包括支持向量机 [3] ，贝叶斯网络 [4]和深度学习 [5]等算法在各个级别上进行了几种分析研究，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。

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。

2 RELATED WORK

我们的课题研究总体可以分为两个部分，新闻情绪分析以及股票预测。情绪分析作为一种 NLP 技术，主要用于挖掘和评估文本/演讲中表达的意见 [6]，目前针对情绪分析的研究也越来越多，主要是看中其潜在的应用价值 [7]。我们这里主要针对新闻中包含的情绪来预测对股票价格的影响。小部分文章理解新闻文章背后的情感是直接将整条新闻输入神经网络进行训练，大多数的方法是识别新闻文章中的重要词语及其极性。如考虑到一个单词可能既可能在利好新闻中出现，也可能在利空新闻中出现，M. I.Yasef Kaya 等人选择以“名词+动词”这种短语的形式输入网络训练得到单词极性 [8]。张泽亚等人则是提出了一种基于单词点互信息的新闻特征抽取方法，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 [9]。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 [10]。而梁士利 [11]等人利用 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。

在股票预测方面，Saloni Mohan 等人将当前新闻的极性和时间 t 之前的股票价格 Price_(t-1)来训练 RNN 模型 [10]。贝勒大学的赵磊 [12]提出了一种离群值挖掘算法，根

据股市高频滴答声数据的数量序列来检测异常，使用交易量分布异常来预测股票价格的上升趋势。

3 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 [13]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。

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

4 PROGRESS

4.1 Preprocessing

我们的处理过程如图 1 所示：

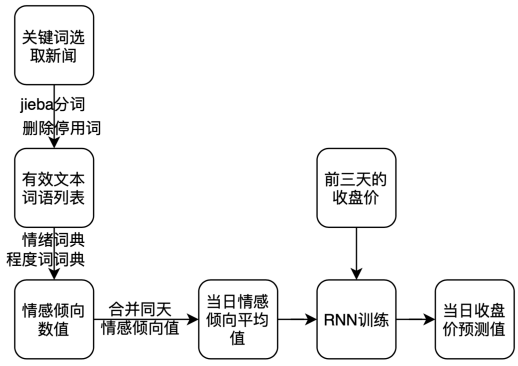


Fig1: 预处理、训练流程

首先，我们读入 2020 年 5 月到 2020 年 9 月的来自东方财富与新浪财经的新闻 csv 文件，共 141423 条新闻。每条新闻都有五个特征，包括新闻 ID、时间、来源、标题和内容，如图 2 所示。根据标题中的关键字“药”，我们获得了 7517 条与医药行业相关的新闻并提取出新闻内容。利用 jieba 分词工具对新闻内容进行分词处理，由于新闻内容中存在很多与情感倾向无关的词语，所以我们再根据哈工大停用词词典去除掉文本中的停用词。

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……

Fig2: 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。

接下来，我们自己建立了与医药行业财经新闻相关的积极情感词典与消极情感词典，找到新闻文本中的积极词和消极词分析情感倾向。在用 jieba 分词之后我们发现对于类似像“销售”这样的单词，如果出现在“某公司销售增加百分之多少”的新闻中是一个积极词性的单词，但如果出现在“某公司销售降低多少百分点”这样的新闻中又是一个表示消极词性的单词。不同于 Yasef Kaya 等人选择将新闻分为一个个“名词+动词”短语输入网络训练 [8]，我们选择了一个现有的 NTUSD-Fin 词典，在这个词典中考虑到类似“销售”这样的词语存在的模棱两可性，该词典已经根据大量的新闻得到了每一个单词的最终值，可以理解为每一个词语积极消极的抵消值 [14]。我们选取 NTUSD-Fin 中词语极性值大于 0 的值为积极词语，小于 0 的词语为消极词语。再根据我们研究新闻的行业特点建立了我们的情感词典。

接着我们利用知网情感分析用词语集 beta 版 [15]中的程度词词典和否定词词典对积极词和消极词加权，最终得到整篇新闻的积极倾向值与消极倾向值。最后把同一天中的新闻合并，得到这一天中所有相关新闻的积极倾向平均值与消极倾向平均值。

情感极性	数量/个	示例
most	69	非常、极端、绝等
very	42	多么、格外、很等
more	37	更加、较、还要等
ish	29	略微、一点、一些等
insufficiently	12	不怎么、轻度、不大等
pos	146	上涨、增加、疫苗等
neg	168	损失、下滑、减少等
inverse	58	不、别、没有等

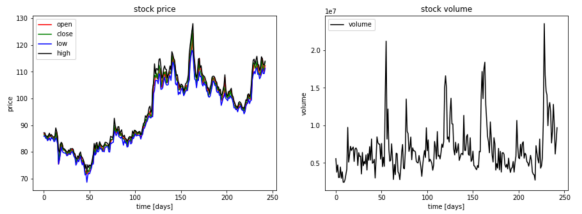
Fig3: 情感词典事例。

通过我们的算法，从单只股票新闻关键词（例如“云南白药”）我们提取出的新闻 42 条，其涵盖到了 25 天。由于新闻数据较少，对后续模型训练的效果不明显，我们将关键词从限定在某一支股票（如“云南白药”）扩充成限定在某一个行业（如药业）。

4.2 Training

对于某只药业股票，我们从聚宽量化交易平台上提取了其从 2020 年 1 月 1 日至 12 月 31 日的股价，原有股票数据有 7 个特征（包括日期 date、开盘价 open、收盘价 close 等等）；从得到的药业相关的 2020 年 129 条新闻，新闻数据有 3 个特征（日期 date、正向情感程度 pos、负向情感程度 neg）。首先我们将处理好的新闻数据和股票数据按照时间整合起来，得到情感分析与新闻数据结合数据集，其包含包括日期 date、收盘价 close、正向情感程度 pos、负向情感程度 neg 等 9 个特征，结果如图 4 所示。

	date	open	close	high	low	volume	money
0	2020-01-02	87.03	86.37	87.13	85.83	5570437.0	4.804153e+08
1	2020-01-03	86.41	86.16	87.13	85.95	3771176.0	3.262361e+08
2	2020-01-06	85.55	85.16	86.07	85.10	4734226.0	4.046714e+08
3	2020-01-07	85.04	85.64	85.95	85.03	3061176.0	2.618692e+08
4	2020-01-08	85.58	84.64	85.58	84.19	3085713.0	2.617135e+08



	date	pos	neg	open	close	high	low	volume	money
0	2020-05-06	39.312500	27.687500	87.03	88.29	88.67	86.08	8919331.0	7.801134e+08
1	2020-05-07	27.322581	34.516129	87.83	87.47	87.99	86.84	6539936.0	5.703585e+08
2	2020-05-08	118.615385	10.923077	87.86	88.47	89.41	87.58	7096194.0	6.294167e+08
3	2020-05-11	175.387097	13.959677	88.58	86.26	89.45	85.68	8450599.0	7.337356e+08
4	2020-05-12	49748.840000	-469770.640000	86.26	86.18	86.83	85.56	5431176.0	4.674289e+08

Fig4: 上图为股票数据，中图为股票数据特征（左图为开盘价、收盘价、最高价、最低价，右图为交易量）随天数的变化趋势，下图为股票数据和新闻情感合成后的数据。

我们将新的数据集绘制出一个相关性矩阵，以查看各特征之间的相关性。图中可以看出，股票特征之间相关性相对较强（open、high 相关性接近 1，open、volume 相关性接近 0.5），而情感特征和股票特征之间的相关性较弱（pos 和 open 相关性 0.1）。由于 pos、neg、close 的变量相关性较弱，它们之间不存在重复特征，在之后的神经网络训练中可以将 pos、high、close 筛选出来都放入训练。

我们将$c_{t-l}, c_{t-l+1}, \ldots, c_{t-1}, p_t, n_t$数据作为输入，来预测输出$c_t$。$c_t$为 t 时间的收盘价，$p_t$为 t 时间的新闻正向情感程度，$n_t$为 t 时间的新闻负向情感程度，l 为滞后参数。模型通过获取 t 天以前 l 天的股票收盘价数据，以及 t 天的新闻情感，来预测 t 天的收盘价。我们设计的 LSTM 网络如图，包括 30651 个参数。我们通过训练 50 个 epoch，设置前 80%数据作为训练数据，后 20%数据作为测试数据。

5 RESULTS

最后的实验结果如图 5，可见我们的预测结果在趋势上与测试数据基本相同。

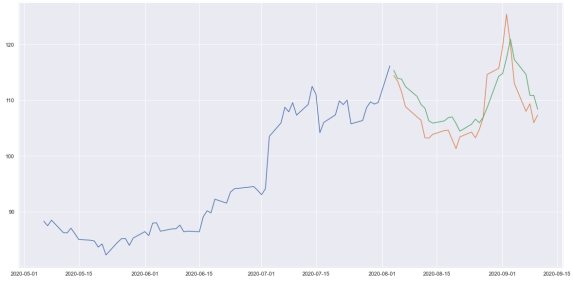


Fig5: 图为股票收盘价预测结果。横轴表示时间，纵轴表示股价。蓝色线表示训练数据集，绿色线表示测试数据预测结果，黄色线表示测试数据。

考虑到新闻情感对实验结果的影响。我们分别进行了仅提取股市收盘价信息，和提取新闻情感及股市收盘价对股票涨跌的影响的实验，统计了测试数据集上的 MAPE，结果如图 6。

	无新闻	有新闻
以岭药业	0.02195	0.01763
复星医药	0.19965	0.10607
恒瑞医药	0.01953	0.01683
片仔癀	0.03936	0.03703
云南白药	0.02769	0.02045

Fig6: 图为对于不同股票仅提取股市收盘价信息，和提取新闻情感及股市收盘价对股票涨跌的影响股票收盘价的 MAPE。

6 ISSUES AND CHALLENGES

在我们的预测方式中，也存在一些问题。根据公司的新闻去预测可能存在的股票涨跌情况，但是事实上可能存在新闻标题是反向预测股票真正的走势，这会给预测结果带来一些偏差。同时也存在有些股市新闻中并不包含任何利空或者利空的关键词，对股票的预测带来了一定难度。另外在大数据、智能化的时代，股民已经不单单从公司新闻这一个渠道

获得消息，手机短视频等也在很大程度上影响了股民的选择[8]。

我们现在使用的上科大云盘中的财经新闻数据只包括2020-05-03至2020-09-10的新闻，在股价预测过程中，我们发现新闻数据量太小，无法对股价进行很好的预测。所以在该项目后面的部分，我们会考虑寻找更大的新闻数据集。

另外我们选用的词典并不是针对医药行业的词典，如果考虑更有针对性的词典我们的准确率可能更高。

7 TIMETABLE AND WORKLOAD DIVISION

任务列表	week1-4	Week5-8	week9-10	Week11-14
阅读文献	沈昭正、井文可、钱丽玥			
讨论实现方法	沈昭正、井文可、钱丽玥			
爬取股票数据及预处理		沈昭正		
筛选新闻数据及预处理		井文可、钱丽玥		
训练		沈昭正、井文可、钱丽玥		
测试			沈昭正、井文可、钱丽玥	
分析实验结果			沈昭正、井文可、钱丽玥	
完成报告、ppt				沈昭正、井文可、钱丽玥

Fig7: 任务列表的分工和完成情况。

REFERENCES

[1] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," Expert Systems with Applications, vol. 38, no. 8, pp. 10389-10397, 2011.Conference Short Name:WOODSTOCK'18

[2] D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

[3] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, vol. 38, no. 3, pp. 2177-2186, 2011.

[4] L. S. Malagrino, N. T. Roman, and A. M. Monteiro, "Forecasting stock market index daily direction: A Bayesia Network approach," Expert Systems with Applications, vol. 105, pp. 11-22, 2018.

[5] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Systems with Applications, vol. 83, pp. 187- 205, 2017.

[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1-135, 2008.

[7] H. Isah, "Social Data Mining for Crime Intelligence: Contributions to Social Data Quality Assessment and Prediction Methods," University of Bradford, 2017.

[8] M.I. Yasef Kaya, Karsl1Gil M E . Stock Price Prediction Using Financial News Articles[C]// 0.

[9] 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

[10] 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.

[11] 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.

[12] ZHAO L, WANG L. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm; proceedings of the Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on, F, 2015 [C]. IEEE.

[13] 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 (February2009),19pages.DOI:https://doi.org/10.1145/1462198.1462204.

[14] Chen C C， Huang H H， Chen H H . NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications.

[15] 董振东.知网[CP/OL]. [2012-03-24]. http://www.keenage.com