山西大学学报(自然科学版)47(4):776—785,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024111

大规模中文金融情感分析数据集自动构建

李大宇[1],李旸[1*],王素格[2]
(1.山西财经大学 金融学院,山西 太原 030006;
2.山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要:金融文本中蕴含着丰富的情感信息,对于捕捉金融市场情绪波动、辅助投资者决策以及实施金融风险管理等都具有重要意义。然而,金融文本的情感标注需要大量领域专家知识,人工标注成本高昂。文章设计了一种基于表情符号远监督指导的自动标注策略,将金融文本中表情符号所表达的情感含义作为指导信息,自动标注文本的情感倾向,构建基础标注数据集;在此基础上,采用持续学习算法训练金融文本情感分类器,预测未标注数据,生成伪标签样本,进一步扩展标注数据集。最终,自动构建了一个包含923万余条股市评论的大规模中文金融情感分析数据集StockSentCN。在人工评估体系下,该数据集的Kappa一致性系数达到0.85,加权平均$F1$值达到90.34%,证明了所构建数据集的高质量与可靠性。数据集公开下载地址为:https://github.com/lidayuls/StockSentCN/。

关键词:中文金融情感分析;股市情感;数据集构建;表情符号;持续学习
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)04-0776-10

Automatic Construction of Large-scale Chinese Financial Sentiment Analysis Dataset

LI Dayu[1], LI Yang[1*], WANG Suge[2]
(1. *School of Finance, Shanxi University of Finance and Economics, Taiyuan* 030006*, China;*
2. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China*)

Abstract: Rich sentiment information is embedded in financial texts, which is of great significance for capturing fluctuations in financial market sentiment, aiding investor with decision-making, and implementing financial risk management. However, sentiment annotation in financial texts requires extensive domain expertise, making manual annotation costly. This paper designs an automatic annotation strategy based on distant supervision guided by emojis, utilizing the sentiment connotations conveyed by emojis in financial texts to automatically label the sentiment polarity, thereby constructing a foundational labeled dataset. On this basis, the continual learning algorithm is employed to train a financial text sentiment classifier, predicting sentiment for unlabeled data and generating pseudo-labeled samples, and further augmenting the labeled dataset. Ultimately, a large-scale Chinese financial sentiment analysis dataset named StockSentCN, encompassing over 9.23 million stock comments, is automatically constructed. Under the human evaluation system, the Kappa consistency coefficient of the dataset reached 0.85, and the weighted average $F1$ score reached 90.34%, proving the high quality and reliability of the constructed dataset. The dataset is publicly available at: https://github.com/lidayuls/StockSentCN/.

Key words: Chinese financial sentiment analysis; stock market sentiment; dataset construction; emojis; continual learning

收稿日期:2024-03-21;**接受日期**:2024-05-20
基金项目:国家自然科学基金(62306169;62106130);山西省基础研究计划项目(202203021212499;20210302124084);山西省高等学校科技创新项目(2022L271)
作者简介:李大宇(1992-),男,山西晋城人,博士,讲师,研究方向为情感分析。E-mail:lidayu@sxufe.edu.com
* **通信作者**:李旸(LI Yang),E-mail:liyang_sxu@163.com
引文格式:李大宇,李旸,王素格.大规模中文金融情感分析数据集自动构建[J].山西大学学报(自然科学版),2024,47(4):776-785. DOI:10.13451/j.sxu.ns.2024111

0 引言

随着互联网和社交媒体的迅猛发展,金融领域的文本资料呈现出爆炸式的增长。这些海量的金融文本数据中蕴含了丰富的情感信息,能够直接反映出金融市场中各类参与者的情绪波动和观点立场。深入挖掘这些情感信息,对于捕捉市场情绪波动、辅助投资决策以及实施风险管理等方面具有重要意义[1]。对于金融文本情感分析而言,底层标注数据的质量和规模直接影响情感分析模型的性能。因此,面对海量的无标签金融文本数据,如何有效的标注其情感倾向,构建大规模高质量的金融情感分析数据集,是提升金融文本情感分析水平的关键。

现有的金融情感分析数据集构建工作主要集中在对英文新闻或推文的人工标注。Malo等[2]构建了一个英文金融情感分类基准数据集Financial PhraseBank,它包含了4 846条从金融领域新闻中挑选的句子,并经过16位金融领域专家的标注。Financial PhraseBank数据集要求标注者从投资者的角度进行情感标注,即新闻是否可能对股价产生“积极”“消极”或“中性”的影响。Cortis等[3]在国际语义评测SemEval 2017中提出了一项金融推文细粒度情感分析任务,所提供的数据集汇聚了2 510条来自Twitter和StockTwits的推文。不同于之前“积极”“消极”以及“中性”的离散型情感标注体系,标注者为每条推文进行打分,并标注一个介于−1(负面)到1(正面)之间的浮点值,0代表“中性”情感。Maia等[4]构建了一个属性级的金融情感分析数据集FiQA,包含金融领域的英文文本实例,其中包括微博消息774条,新闻标题529条。数据集标注了文本中所讨论的方面、文本中提到的对象以及情感分数。其情感分数同样使用从−1(负面)到1(正面)的连续数值来定义。Xing等[5]构建了StockSen数据集,由2019年6月至8月期间从StockTwits平台提取的20 675条金融推文组成,所有推文都被标注了“正面”或“负面”的情感标签。近来,Pei等[6]人工构建了一个由2 113条推文组成的股市金融情感分析数据集TweetFinSent,他们基于用户是否预期从股票投资中获得正面或负面回报来标注推文的情感倾向,为研究股票市场情绪提供了新的视角。尽管上述数据集被广泛应用于评测金融情感分析任务[7-9],但普遍存在以下问题:1)数据集的标注需要大量金融领域专家,人工标注成本高昂;2)数据规模较小,难以满足大模型训练对数据量的需求;3)大部分标注数据集来自于英文语料,公开的中文金融情感分析数据集数量较少,阻碍了中文金融情感分析社区的发展。

为了解决这些问题,本文提出了一种利用表情符号作为远监督指导信息的大规模中文金融情感分析数据集自动构建方法。如表1所示,股民在股市论坛发表评论时倾向于在文本内容中增加表情符号来增强自己的情感表达。例如,在看涨股市的评论中,股民可能会添加积极的表情符号,如“📈”“🧧”“👍”;在看跌股市的评论中,股民可能会添加消极的表情符号,如“📉”“🍉”“😱”。这些表情符号可以作为投资者观点和情感倾向的指示信息,充分利用这些表情符号所表达的情感含义作为远监督指导信息,可以大大减少人工标注成本。

具体而言,本文以中国股票市场为例,系统地爬取了东方财富股吧论坛1990年至2023年间881家上市公司9 111余万条原始评论数据,并进行数据清洗与预处理。接着,提取原始评论中含有表情符号的评论文本,根据其表情符号所表达的“看涨”与“看跌”的情感含义,自动

表1 带有表情符号的股市评论示例

Table 1 Examples of stock market comments with emojis

Post_id	Bar_name	Post	Sentiment
1386064352	平安银行	也来参加一份,等等涨停,明天再涨停就快回本了🍌🍌📈📈📈📈📈📈	Positive
815694247	格力电器	今早58元全仓追进🧧🧧👍	Positive
1321497498	科大讯飞	接下来会是十连阴📉	Negative
1261379304	上海能源	老散们别再做梦了~今年没行情~早点结账回家过年~拜拜📉🍉😱	Negative

标注文本的正负情感倾向,以此构建基础标注数据集。在此基础上,采用持续学习算法训练金融文本情感分类器,并对大量无标签样本进行预测,生成伪标签数据集,以此扩展标注数据集。

本文的主要贡献如下:

(1)提出了一种大规模金融文本情感分析数据集自动构建方法,利用表情符号作为远监督指导信息,自动标注情感标签;在此基础上,提出了一种基于持续学习的伪标签样本获取方法,进一步扩展标注数据集。

(2)构建了一个大规模的中国股票市场金融文本情感分析数据集StockSentCN。数据集来自“平安银行”“万科”“中国宝安”“深科技”等在内的881家上市公司股吧论坛,广泛涉及科技、消费、能源、地产等多个行业,包含923万余条股市评论文本。

1 数据集构建

1.1 数据采集

首先进行数据采集,数据来源为中国知名财经社区——东方财富股吧网(http://guba.eastmoney.com)。使用Python爬虫工具库,以自动化的方式收集评论文本数据,并对原始文本进行初步的数据预处理和清洗。最终搜集到1990年01月至2023年12月间发布的9 111万余条股吧评论文本数据,涵盖了包括“平安银行”“万科”“中国宝安”“深科技”等在内的881家上市公司股吧论坛。图1展示了最终爬取数据的样例。

1.2 数据标注

(1)金融情感类别标签定义

不同于传统情感分析关注于用户对产品或服务的个人喜好、情绪表达和主观评价[10],金融领域的情感分析更侧重于分析投资者对市场动态、个股表现或宏观经济的观点与预期[11-12]。针对股票市场的情感分析,研究者更多地关注于投资者对股票价格走势的预期是积极(看涨)的还是消极(看跌)的,而非关注个人的内在情绪或偏好。因此,我们定义股市评论文本的金融情感类别标签含义如下:

消极(看跌):看跌股票市场,认为股票的价格或价值将会下降;

积极(看涨):看涨股票市场,认为股票的价格或价值将会上升;

中性:未明显表达消极(看跌)或积极(看涨)的情感倾向。

(2)表情符号远监督指导

表情符号广泛存在于文本内容中,并逐渐成为社交媒体和论坛中不可或缺的交流元素[13]。它通过视觉符号的方式传达人类的面部表情、手势和动作等,弥补了纯文本内容情感表达的不足,具有增强或改变文本情感色彩

post_id	post	stockbar_name	post_publish_time	user_nickname	post_click_count	post_comment_count
946875870	明天可以逢低买入。	平安银行吧	2020-07-15 17:34:47	小小技术员	482	0
946869030	历史惊人的相似，调整十日线附近，银行股会拉	平安银行吧	2020-07-15 17:07:17	欢乐新豆豆	754	0
946868300	最近几个月的操作已经很有耐心了，也怪自己賬	平安银行吧	2020-07-15 17:04:13	朱小肉	769	2
946864351	了，银行真是天天有利空，今天是深圳房地产调	平安银行吧	2020-07-15 16:48:08	芨芨菜	528	0
946864034	这么狂出，估计没几天就13了	平安银行吧	2020-07-15 16:46:52	鱼跃龙门3734	692	1
946860640	【0715收市点评】有望探底回升	财富号评论吧	2020-07-15 16:33:32	财通证券资管	7513	1
946850947	弘缠院：失守3400！A股又大跌！连续调整为哪	财富号评论吧	2020-07-15 15:57:29	弘缠院	10065	1
946849292	关灯吃面，再见	平安银行吧	2020-07-15 15:51:36	mme185349	410	0
946848109	这几天不把散户折腾死，狗主力是没完的。	平安银行吧	2020-07-15 15:47:54	稳妥中有赢	431	0
946844542	未来2个月银行股有路吗	平安银行吧	2020-07-15 15:37:27	股友veY0T1	2036	8
946841761	你们在经历牛市，我在经历股灾[哭]	平安银行吧	2020-07-15 15:29:43	玉立骄阳	675	1
946841128	上涨都是利好，下跌都是利空，平安你就是个朱	平安银行吧	2020-07-15 15:28:05	大凡至简	445	0
946840387	差不多，随它砸，不动了	平安银行吧	2020-07-15 15:26:26	玉立骄阳	539	2
946839903	这股我赚两千觉得少，没卖，现在赔钱。我是该	平安银行吧	2020-07-15 15:24:46	xw红红	778	7
946839494	贷款用途造假，平安银行被罚90万元	平安银行吧	2020-07-15 15:23:59	钱亿丰盛	548	1
946834914	扶我起来，我还能补仓！！[哭][哭][哭]	平安银行吧	2020-07-15 15:14:06	蜡笔小葵	394	0
946833888	庄家吓散户，尾盘砸盘的都是小心眼的庄家，医	平安银行吧	2020-07-15 15:12:19	大山深处老农民	468	0
946833657	还要跌完明天	平安银行吧	2020-07-15 15:11:59	海纳百川OK	358	0
946833535	精准抄顶，我也服了我自己。	平安银行吧	2020-07-15 15:11:35	四拾三拱原本出	550	1
946832992	贷款用途造假 平安银行被罚90万元	平安银行吧	2020-07-15 15:10:16	平安银行资讯	5487	13
946829207	死股，太可恨了，害老子亏惨了	平安银行吧	2020-07-15 15:04:52	n2522132332198 28	378	0
946828187	今天一直补，一直补，还没看到头[哭][哭][哭	平安银行吧	2020-07-15 15:03:37	新鲜的小韭菜88	565	1
946825259	庄家吓散户，尾盘砸盘的都是小心眼的庄家，医	平安银行吧	2020-07-15 14:59:51	大山深处老农民	379	0
946823011	不到13.50左右看来都止不住。	平安银行吧	2020-07-15 14:57:10	两棵树	485	1
946822347	快跌吧，哥等着加仓呢	平安银行吧	2020-07-15 14:56:27	还有八十个秋	336	0
946821690	分红2毛。平安还要暴跌。看10元。	平安银行吧	2020-07-15 14:55:54	股友cVxPxX6291	543	2
946820404	割了，大家留下的发财	平安银行吧	2020-07-15 14:54:32	发股神888888	357	0
946819566	缺口已补，明天正式冲	平安银行吧	2020-07-15 14:53:48	丨一步一脚印丨	363	0

图1 爬取数据示例

Fig. 1 Examples of crawled data

的作用[14-15]。在股市论坛中,表情符号同样被用来表达投资者对股市的情感倾向。股吧论坛表情符号体系中共含有95种表情,我们对原始文本中包含表情符号的文本进行了统计,其中含有表情符号的评论文本大约占总评论数据集的6.76%。本文将表情符号作为远监督指导信息,自动标注评论的情感极性,构建基础标注数据集,具体流程如图2所示。

表情符号筛选 首先,从股吧评论的所有表情符号中筛选明确表达投资者“看涨”和“看跌”股市情感的表情符号。我们从95种表情符号中筛选了7种明确表达“看涨”情感的表情符号:“[看多]”“[上涨]”“[买入]”“[满仓]”“[加仓]”“[抄底]”“[梭哈]”,以及6种明确表达“看跌”情感的表情符号:“[看空]”“[下跌]”“[卖出]”“[空仓]”“[减仓]”“[泡沫]”。表情符号的选择基于其在金融评论文本中频繁出现且表现出明显的“看涨”与“看跌”正负情感倾向。

数据自动标注 将“看涨”与“看跌”类别中的表情符号作为情感指示器,从原始数据集中筛选出包含这些表情符号的评论文本,并根据表情符号的类别将其情感倾向分别标注为“积极”与“消极”。在自动标注过程中,剔除那些同时包含“看涨”与“看跌”两类极性相反表情符号的评论文本,确保文本所表达的情感极性只属于“积极”与“消极”中的某一类情感类别。同时,为了防止在未来的数据训练过程中,模型过度依赖表情符号作为文本情感识别的关键特征而忽略文本内容本身,我们在标注情感标签之后,移除了文本中的表情符号,仅保留了文本信息。这样做的目的是确保模型能够更准确地从文本内容中学习到情感特征。

数据筛选 经过上述步骤后,得到的标注数据中仍可能存在少量标注错误。例如,由于表情符号使用不规范导致的表情符号与文本内容所表达的情感倾向不符。因此,还需对标注数据进一步筛选和校对。我们将自动标注的样本均分为5份,在文本情感分类器上进行五折交叉验证,以此剔除不一致样本。具体为,首先,将其中4份作为训练集去训练文本情感分类器;其次,使用训练好的文本情感分类器预测第5份数据,得到预测结果,并从标注数据中剔除预测结果与自动标注结果不一致的样本;最后,切换数据组合,依次进行5次独立的模型训练和验证。此外,经过多轮五折交叉验证筛选后,我们还将剩余标注样本按照模型预测的置信度进行排序,进一步剔除置信度较低的样本。

“中性”情感类别标注 由于上述构建的标注数据集中只包含“积极”和“消极”两类情感,因此还需要人工标注“中性”样本。我们从原始数据中随机抽取了部分评论文本,将其中没有明显表达消极(看跌)或积极(看涨)情感倾向的评论文本标注为“中性”,人工构建了少量“中性”类别样本。

(3)伪标签数据扩展

在上述基础数据集构建完成后,我们提出了一种基于持续学习的伪标签数据扩展方法来进一步扩充标注数据集。其流程图如图3所示。首先,使用初始带标签基础数据集L_0来对

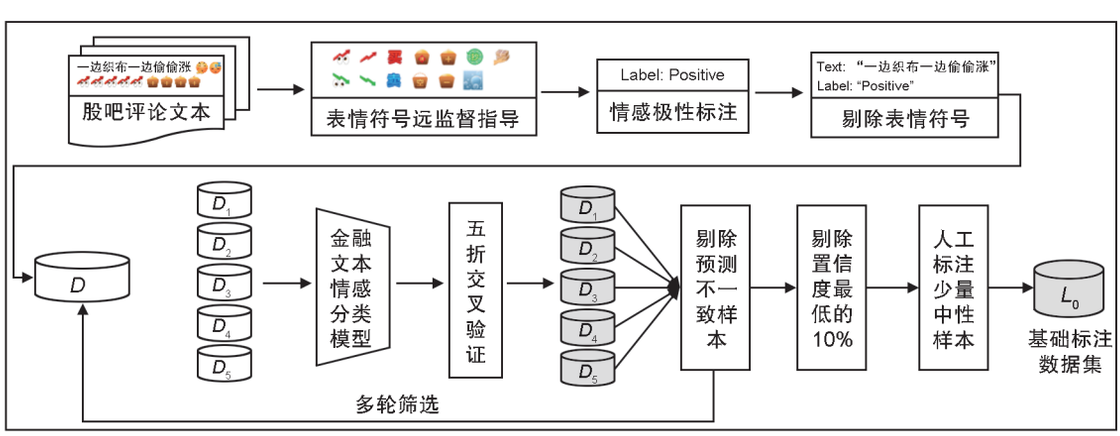

图2 基于表情符号远监督指导的数据标注流程图

Fig. 2 Data labeling flowchart guided by emoji-based distant supervision

初始的文本分类模型M进行训练。然后,将所有无标签数据U按照所在个股将其分为多个子集,即$U=\{U_1,U_2,\cdots,U_n\}$,模型M在每个子集上分批次进行持续学习训练。对于子集U_i中的每个样本x,使用模型M预测得到样本标签的概率分布$p(y|x)$,并计算这个预测概率分布的熵$Entropy(p(y|x))$。此熵值反映了预测结果的不确定性,如果这个熵值小于预先设定的置信度阈值θ,说明模型认为其对该样本的预测是可靠的,此时就将预测概率最大的标签y作为样本x的伪标签,构建伪标签样本(x,y)。每个子集U_i中的所有伪标签样本构成了伪标签数据集$L_i=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$。

接下来,使用模型M在新扩展的伪标签数据集$L_i=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$上进行基于持续学习的训练策略,使模型$M$能够不断从新增数据中学习新的知识。本文采用了一种基于知识蒸馏的持续学习策略[16],将模型M分为教师模型M(teacher)和学生模型M'(student),训练过程中将教师模型M作为参照模型,只对学生模型M'进行参数更新,训练结束后使用学生模型M'的参数更新教师模型M。训练损失函数$\mathcal{L}$包括了两个部分:分类任务的交叉熵损失$\mathcal{L}_{CE}$和持续学习的Kullback-Leibler(KL)散度损失$\mathcal{L}_{KL}$,两者的权重由参数λ决定。交叉熵损失$\mathcal{L}_{CE}$确保学生模型M'可以从新增数据中学习新的知识,而KL散度损失$\mathcal{L}_{KL}$则防止学生模型偏离教师模型过远,确保在学习新知识的同时保留已学到的知识,避免了持续学习中的灾难性遗忘问题。基于知识蒸馏的持续学习算法损失函数如下式所示:

$$\mathcal{L}=(1-\lambda)\times\mathcal{L}_{CE}+\lambda\times\mathcal{L}_{KL}, \tag{1}$$

$$\mathcal{L}_{CE}=-\frac{1}{N}\sum_{i=1}^{N}y_i\log(p_i^{student}), \tag{2}$$

$$\mathcal{L}_{KL}=D_{KL}(p^{teacher}\|p^{student})=\frac{1}{N}\sum_{i=1}^{N}p_i^{teacher}\log(\frac{p_i^{teacher}}{p_i^{student}}), \tag{3}$$

其中$\mathcal{L}$为基于知识蒸馏的持续学习算法损失函数,$\mathcal{L}_{CE}$为交叉熵损失,$\mathcal{L}_{KL}$为KL散度损失,λ为持续学习中蒸馏损失的权重。y是真实的标签分布,i表示第i个样本。$p_i^{teacher}$是教师模型预测样本i的概率分布,$p_i^{student}$是学生模型预测样本i的概率分布。

1.3 数据统计

本节对所构建的大规模中文金融文本情感分析数据集StockSentCN进行了统计,如表2所示。其中,通过远监督指导方法标注的数据总计191 129条,消极、积极和中性评论数量分别为50 261、126 903和13 965;通过基于持续学习的伪标签扩展方法标注的数据总计9 047 422条,消极、积极和中性评论分别为2 645 371、4 836 677和1 565 374。最终,StockSentCN数据集的消极、积极和中性评论总数分别为

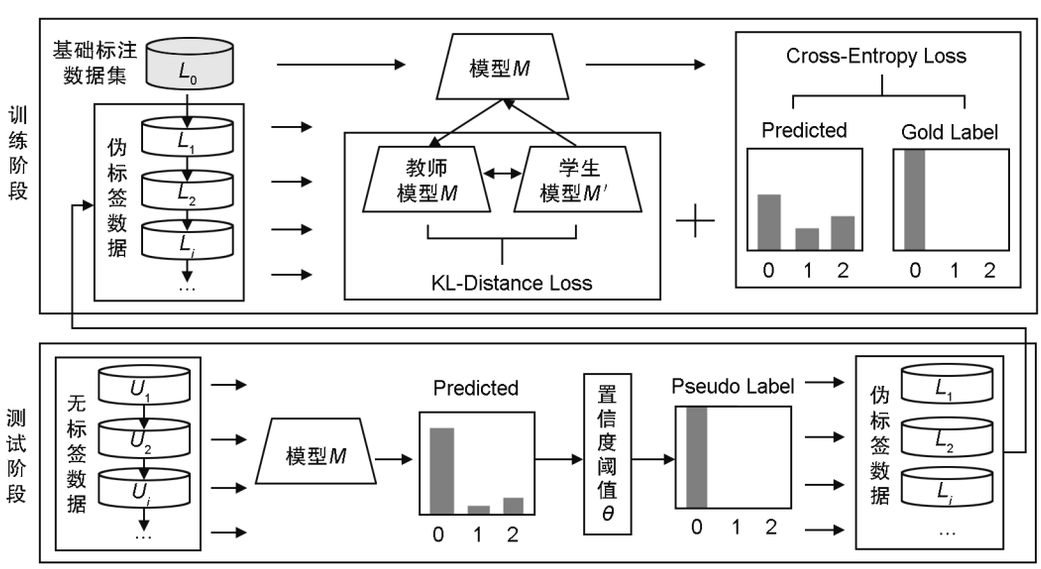

图3 基于持续学习的伪标签数据扩展流程图

Fig. 3 Flowchart of continual learning-based pseudo-label data augmentation

2 695 632、4 963 580和1 579 339,总计9 238 551条。本文的数据集涵盖了从1990年至2023年间多个行业的股市评论,确保了数据的广泛性和代表性。

表2 StockSentCN数据集统计信息

Table 2 Statistics of StockSentCN dataset

标注方法	消极	积极	中性	总计
远监督指导	50 261	126 903	13 965	191 129
伪标签扩展	2 645 371	4 836 677	1 565 374	9 047 422
总计	2 695 632	4 963 580	1 579 339	9 238 551

我们还对StockSentCN数据集排名前十个股的情感标签分布情况进行了统计,如图4所示。数据总量排名前十的个股为:“京东方A”“格力电器”“中兴通讯”“东方财富”“长安汽车”“比亚迪”“天齐锂业”“罗牛山”“三六零”“欧菲光”。

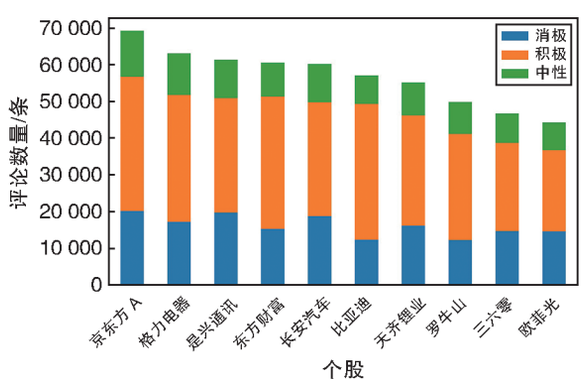

图4 StockSentCN数据集排名前十个股的情感标签分布

Fig. 4 Sentiment label distribution for the top ten ranked stocks in the StockSentCN dataset

1.4 数据集评估

本节设计实验来人工评估StockSentCN数据集的标注质量。首先,从StockSentCN数据集中进行随机抽样,共随机抽取3组数据,每组300个样本。接着,对3位金融领域专家进行情感标注培训,每位专家独立标注1组样本。在获取专家人工标注的标签后,通过计算Cohen's Kappa值[17]来评估人工标注标签与自动标注标签之间的一致性。此外,将人工标注的标签视为“黄金标准”标签,将StockSentCN自动标注标签作为预测标签,采用分类任务中的准确率、宏评价$F1$值以及加权平均$F1$值等指标来衡量自动标注标签的准确性。

最终,三组样本的平均Kappa系数为0.85,证明了人工专家标注标签与StockSentCN数据集自动标注标签之间存在较高的一致性[18]。三组样本的平均准确率为90.66%、宏评价$F1$值为88.45%、加权平均$F1$值为90.34%,这些分类指标同样说明了人工专家标注标签与StockSentCN自动标注标签之间的差异较小。

综上所述,通过从StockSentCN自动标注数据集中随机抽样,并获取领域专家人工标注结果,采用Kappa系数与分类指标来评价人工专家标注标签与StockSentCN自动标注标签之间差异,结果表明了StockSentCN数据集具有较高的自动标注质量。

2 实验

本章在所构建数据集的上对比了多种基线文本分类模型的性能,重点考察神经网络模型和预训练语言模型在所构建数据集上的表现。

2.1 数据切分与评价指标

我们在基础数据集上进行了实验,数据集按照8:1:1切分为训练集、验证集、测试集。详细的数据统计如表3所示。分类评价指标采用准确率、宏平均$F1$值以及加权平均$F1$值。

表3 实验数据集切分情况

Table 3 Segmentation of experimental dataset

	消极	积极	中性	总计
训练集	40 255	101 428	11 220	152 903
验证集	5 052	12 696	1 365	19 113
测试集	4 954	12 779	1 380	19 113
总计	152 903	19 113	19 113	191 129

2.2 基线方法

我们对比了三类模型在所构建数据集上的性能:传统机器学习模型、基础神经网络模型以及大规模预训练语言模型。传统机器学习模型包括K最近邻(k-Nearest Neighbor,KNN)、多项式朴素贝叶斯(Multinomial Naive Bayes,MultinomialNB)、决策树(Decision Tree)、伯努利朴素贝叶斯(Bernoulli Naive Bayes,BernoulliNB)、随机森林(Random Forest)、逻辑回归(Logistic Regression)、支持向量机(Support Vector Machine,SVM)等;基础神经网络模型包括卷积神经网络(Convolutional Neural Network,CNN)、双向长短期记忆网络(Bidirectional Long Short-

Term Memory,Bi-LSTM)、带注意力机制的双向长短时记忆网络(Bidirectional Long Short-Term Memory with Attention, Bi-LSTM+ATT)、双向门控循环单元(Bidirectional Gated Recurrent Unit,Bi-GRU)和带注意力机制的双向门控循环单元(Bidirectional Gated Recurrent Unit with Attention,Bi-GRU+ATT)等(ATT为注意力机制);大规模预训练模型包括文本到文本转换器(Text-To-Text Transfer Transformer,T5)、生成式预训练转换器(Generative Pre-trained Transformer,GPT)、双向编码器表示转换器(Bidirectional Encoder Representations from Transformers,BERT)、知识增强表示(Enhanced Representation through Knowledge Integration,ERNIE)以及鲁棒优化的BERT预训练方法(Robustly Optimized BERT Pretraining Approach,RoBERTa)等。对于大规模预训练模型,使用了不同参数量和基于不同预训练数据的多个版本。例如,T5模型[19]包括"T5 small model"和"T5 base model"两个版本,GPT模型[20]包括"GPT-2 model"和"GPT-2 model for Chinese"两个版本,BERT模型[21]包括"Distilled BERT base model"、"BERT base model"、"BERT base model for Chinese"和"Chinese BERT with Whole Word Masking"四个版本,ERNIE模型[22]包括"ERNIE 3.0 Nano Chinese model"、"ERNIE 3.0 Mini Chinese model"和"ERNIE 3.0 Base Chinese model"三个版本,RoBERTa模型[23]包括"RoBERTa base model"和"Chinese RoBERTa with Whole Word Masking"两个版本。

2.3 实验结果

表4展示了传统机器学习模型、基础神经网络模型以及大规模预训练模型三类模型在基础标注数据集上的情感分类结果。从表中可以看出,传统机器学习模型总体表现在75%~85%(加权平均$F1$值)之间,其中线性核SVM模型加权平均$F1$值达到最优的85.19%;基础神经网络模型如CNN、Bi-LSTM、Bi-LSTM+ATT、Bi-GRU和Bi-GRU+ATT要优于传统机器学习模型,总体加权平均$F1$值在92%~93%之间,其中Bi-GRU表现最佳,加权平均$F1$值为93.11%。基于大规模预训练模型表现差异性较大,加权平均$F1$值在53%~98%之间。其中"Chinese RoBERTa with Whole Word Masking"版本的RoBERTa模型取得了最优的表现,加权平均$F1$值为98.41%。

对实验结果进行分析可以得出以下结论:1)总体上,基于大规模预训练模型的方法表现最优,基础神经网络模型表现次之,传统机器学习模型表现最差。基础神经网络模型相较于传统机器学习模型,其模型结构能够更有效的建模句子中词与词之间的上下文关系,并且能够捕捉到大规模文本数据中复杂的非线性关系。而大规模预训练模型则是在海量的数据上进行了预训练,这使得它们能够学习到更为丰富的语言表示和深层的语义信息,在处理情感分类任务时表现更加出色;2)大模型预训练模型中,拥有更大参数量以及在更大规模数据集上进行预训练的模型通常表现更好。例如,ERNIE模型三种版本(Nano,Mini,Base)的结果随着模型参数量依次上升(91.39%,94.11%,96.12%);3)预训练模型的中文版本在数据集上表现要优于通用版本。例如,"Chinese BERT with Whole Word Masking"和"Chinese RoBERTa with Whole Word Masking"模型针对中文的特性,使用Whole Word Masking策略,能够更有效地保留中文词汇信息,对于中文文本的语义建模更加准确。

此外,从上述结果还可以看出,各类文本分类模型在我们所构建的大规模中文金融情感数据集上取得了较优的结果,未经过预训练的基础神经网络模型能够取得92%以上的准确率,而中文大规模预训练模型则能达到96%以上的准确率。这一方面证明了模型的有效性,另一方面也说明了我们构建的数据集的质量。我们所构建的大规模中文金融情感数据集包含了关于大量上市公司股市的相关评论,保证了数据的多样性和覆盖面,以便模型能够学习到不同情境下的情感表达和金融术语。数据能够为各类文本分类模型提供足够的训练样本以及丰富的语义信息,从而帮助模型更好地学习和理解金融领域的特定知识和情感倾向。

2.4 持续学习训练参数分析

在第1.2节中,我们使用了基于持续学习的

表4 传统机器学习模型、基础神经网络模型以及大规模预训练模型在基础标注数据集上的情感分类结果

Table 4 Sentiment classification results of traditional machine learning models, basic neural network models, and large-scale pre-trained models on the basic labeled datasets

模型			单类$F1$值/%			总体结果/%		
			消极	积极	中性	准确率	宏评价$F1$值	加权平均$F1$值
传统机器学习模型		KNN	63.91	86.47	14.35	78.47	54.91	75.42
		MultinomialNB	75.56	89.63	2.43	83.27	55.87	79.69
		DecisionTree	72.86	88.85	29.60	81.29	63.77	80.43
		BernoulliNB	78.50	91.23	23.89	84.93	64.54	83.07
		RandomForest	77.95	91.03	28.45	84.78	65.81	83.12
		LogisticRegression	80.44	92.09	35.14	86.35	69.22	84.96
		SVM	80.73	92.11	37.15	86.42	70.00	85.19
基础神经网络模型		CNN	90.71	96.24	59.88	92.52	82.28	92.18
		Bi-LSTM	91.16	96.51	61.26	92.47	82.98	92.58
		Bi-GRU	91.73	96.74	64.46	92.82	84.31	93.11
		Bi-LSTM+ATT	91.07	96.62	61.98	92.55	83.22	92.68
		Bi-GRU+ATT	91.14	96.73	63.18	92.73	83.68	92.86
大规模预训练模型	T5	T5 small model	0.00	80.14	0.00	66.86	26.71	53.58
		T5 base model	0.00	80.11	13.07	66.88	31.06	54.50
	GPT	GPT-2 model	83.50	93.83	39.07	88.06	72.13	87.20
		GPT-2 model for Chinese	95.64	98.52	74.79	96.11	89.65	96.06
	BERT	Distilled BERT base model	51.64	83.77	33.85	74.48	56.42	71.84
		BERT base model	51.15	84.08	32.50	74.71	55.91	71.82
		BERT base model for Chinese	96.80	98.80	78.43	96.84	91.34	96.81
		Chinese BERT with Whole Word Masking	97.24	98.99	81.19	97.27	92.47	97.25
	ERNIE	ERNIE 3.0 Nano Chinese model	89.42	95.86	57.05	91.74	80.78	91.39
		ERNIE 3.0 Mini Chinese model	93.17	97.34	67.63	94.30	86.04	94.11
		ERNIE 3.0 Base Chinese model	96.04	98.55	73.97	96.23	89.52	96.12
	RoBERTa	RoBERTa base model	85.07	93.97	44.61	88.84	74.55	88.10
		Chinese RoBERTa with Whole Word Masking	98.28	99.48	88.94	98.42	95.57	98.41

伪标签数据集扩展方法,并使用了参数λ来确定持续学习中的蒸馏损失权重。在本节中,我们通过使用不同的λ值来观察持续学习方法的有效性以及参数λ对持续学习效果的影响。基础分类模型使用Bi-GRU模型,结果如图5所示。

在还未进行持续学习时,Bi-GRU模型在基础数据集上的准确率为92.82%。当λ=0时,即只使用交叉熵分类损失对模型进行持续学习训练而不添加蒸馏损失,模型的效果会随着来自不同个股数据的增多而快速下降。例如,当个股数达到50时,模型经过持续学习后准确率从92.82%下降为87.28%。这说明模型在进行持续学习训练时,存在灾难性遗忘问题,模型参数可能会过度拟合新数据,从而损害其在旧数据上的表现。从λ取0.3、0.5、0.7以及0.9时模型的表现可以看出,λ取值越小,持续学习中的蒸馏损失权重越小,更新后的模型偏离原模型越远,在原数据的表现也就越差。从结果上

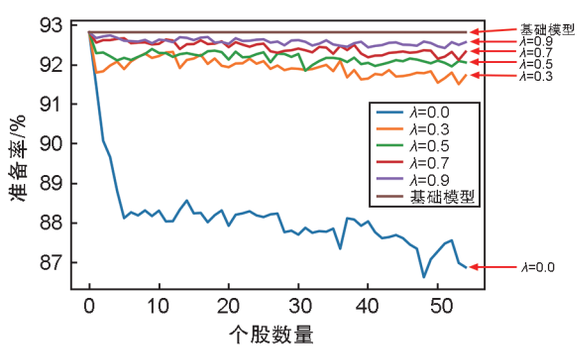

图5 持续学习策略中λ不同取值的实验结果

Fig. 5 Experimental results of different λ values in continual learning strategy

看,当λ=0.9,个股数达到50时,模型准确率从92.82%下降为92.47%,下降幅度较小。这说明在持续学习中增加蒸馏损失的比重可以有效的缓解模型的灾难性遗忘问题。在我们的数据集构建中,λ取值为0.9。

2.5 案例研究

通过进一步分析数据集的样例我们发现,股市评论中存在大量隐式情感现象,这给股市金融情感分析带来了挑战。例如,“我左眼皮在跳”“顶着! 天塌下来,当被盖!”,这些评论中都没有出现明确表示情感倾向的词语描述,但是都表达出了“积极”的股市金融情感。“我左眼皮在跳”的隐含情感解释来自于中国民间的一句俗语“左眼跳财,右眼跳灾”,当描述“左眼皮在跳”时,投资者通常认为预示着好运,预期股市会出现上涨的趋势。“顶着! 天塌下来,当被盖!”这句话中的“顶着”意味着投资者会顶住当前压力,继续持有现在的股票,即使当前面临一些困难和压力,但相信股价未来会继续上涨;而“天塌下来,当被盖”则表示投资者对市场的极度信心,即使出现最坏的情况,他们也相信自己能够应对并从中获益。

此外,评论中反讽、比喻、夸张等修辞手法的运用也增加金融情感分析的难度。例如,“15.489的本啊亏了三分之一了,跌的太慢了,平安银行加油!”这句话使用了反讽的修辞手法,说话者实际上是在表达对平安银行股票跌幅过大的不满(“亏了三分之一”),但是使用了“跌的太慢了”“加油”等词,通常“加油”是用来鼓励和支持的,这里却被用来表达相反的情感。“神精病买了这个病猫”这句话使用了比喻的修辞手法,将投资者比喻为“神精病”,将所购买股票比喻为“病猫”,描述了投资者行为的不理智以及表达了对股票价格的不满。

3 结论

本文针对现有研究缺乏大规模中文金融情感分析公开数据集的问题,构建了一个包含923万余条股市评论的大规模中国股票市场文本情感分析数据集StockSentCN。数据来自包括“平安银行”“万科”“中国宝安”“深科技”等在内的881家上市公司股吧论坛,涉及科技、消费、能源、地产等多个行业。本文首先利用表情符号作为远监督指导信息构建初始基础数据集,接着提出了一种基于持续学习的伪标签样本获取策略来进一步扩展数据集。在多个基准文本分类模型上对该数据集进行了实验和对比分析,验证了数据的质量以及构建方法的有效性。本文所构建数据有望为中文金融情感分析领域的研究和应用提供有力的数据支持。

参考文献:

[1] CHEN C C, HUANG H H, CHEN H H. A Research Agenda for Financial Opinion Mining[J]. *Proc Int AAAI Conf Web Soc Medium*, 2021, **15**: 1059-1063. DOI: 10.1609/icwsm.v15i1.18130.

[2] MALO P, SINHA A, KORHONEN P, *et al*. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts[J]. *J Assoc Inf Sci Technol*, 2014, **65**(4): 782-796. DOI: 10.1002/asi.23062.

[3] CORTIS K, FREITAS A, DAUDERT T, *et al*. SemEval-2017 Task 5: Fine-grained Sentiment Analysis on Financial Microblogs and News[C]//Proceedings of the 11th International Workshop on Semantic Evaluation. Vancouver, Canada: Association for Computational Linguistics, 2017:519-535. DOI: 10.18653/v1/S17-2089.

[4] MAIA M, HANDSCHUH S, FREITAS A, *et al*. WWW '18 Open Challenge: Financial Opinion Mining and Question Answering[C]//WWW '18: Companion Proceedings of the The Web Conference 2018. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018: 1941-1942. DOI: 10.1145/3184558.3192301.

[5] XING F, MALANDRI L, ZHANG Y, *et al*. Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 978-987. DOI: 10.18653/v1/2020.coling-main.85.

[6] PEI Y L, MBAKWE A, GUPTA A, *et al*. TweetFinSent: A Dataset of Stock Sentiments on Twitter[C]//Proceedings of the 4th Workshop on Financial Technology and Natural Language Processing. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022: 37-47. DOI: 10.18653/v1/2022.finnlp-1.5.

[7] KARANIKOLA A, DAVRAZOS G, LIAPIS C M, *et al*. Financial Sentiment Analysis: Classic Methods Vs. Deep Learning Models[J]. *Intell Decis Technol*, 2023, **17**(4):

893−915. DOI: 10.3233/idt-230478.

[8] ZHANG B Y, YANG H Y, ZHOU T Y, *et al*. Enhancing Financial Sentiment Analysis *via* Retrieval Augmented Large Language Models[C]//Proceedings of the 4th ACM International Conference on AI in Finance. New York, NY, USA: Association for Computing Machinery, 2023: 349−356. DOI: 10.1145/3604237.3626866.

[9] ZHANG B Y, YANG H Y, LIU X Y. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-purpose Large Language Models[J]. *SSRN J*, 2023. DOI: 10.2139/ssrn.4489831.

[10] LIU B. Sentiment Analysis and Opinion Mining[M]. Cham, Switzerland: Springer, 2012.

[11] LIU Z, HUANG D G, HUANG K Y, *et al*. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining[C]//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, 2020: 4513−4519. DOI: 10.24963/ijcai.2020/622.

[12] CHEN C C, HUANG H H, CHEN H H. Financial Opinion Mining[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic & Online: Association for Computational Linguistics, 2021: 7−10. DOI: 10.18653/v1/2021.emnlp-tutorials.2.

[13] FELBO B, MISLOVE A, SØGAARD A, *et al*. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1615−1625. DOI: 10.18653/v1/D17-1169.

[14] ALSHENQEETI H. Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-Semiotic Study[J]. *Adv Lang Lit Stud*, 2016, **7**(6): 56−69. DOI: 10.7575/aiac.alls.v.7n.6p.56.

[15] BOUTET I, LEBLANC M, CHAMBERLAND J A, *et al*. Emojis Influence Emotional Communication, Social Attributions, and Information Processing[J]. *Comput Hum Behav*, 2021, **119**: 106722. DOI: 10.1016/j.chb.2021.106722.

[16] WANG L Y, ZHANG X X, SU H, *et al*. A Comprehensive Survey of Continual Learning: Theory, Method and Application[J]. *IEEE Trans Pattern Anal Mach Intell*, 2024:1−20. DOI: 10.1109/TPAMI.2024.3367329.

[17] COHEN J. A Coefficient of Agreement for Nominal Scales[J]. *Educ Psychol Meas*, 1960, **20**(1): 37−46. DOI: 10.1177/001316446002000104.

[18] MCHUGH M L. Interrater Reliability: The Kappa Statistic[J]. *Biochem Med*, 2012, **22**(3): 276−282. DOI: 10.11613/BM.2012.031.

[19] RAFFEL C, SHAZEER N, ROBERTS A, *et al*. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer[J]. *J Mach Learn Res*, 2020, **21**(140):1−67.

[20] RADFORD A, NARASIMHAN K, SALIMANS T, *et al*. Improving Language Understanding with Unsupervised Learning[R/OL]. Technical Report, OpenAI, 2018. https://openai.com/index/language−unsupervised/.

[21] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171−4186. DOI: 10.18653/v1/N19-1423.

[22] SUN Y, WANG S H, LI Y K, *et al*. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding[J]. *Proc AAAI Conf Artif Intell*, 2020, **34**(5): 8968−8975. DOI: 10.1609/aaai.v34i05.6428.

[23] LIU Y H, OTT M, GOYAL N, *et al*. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL]. arXiv Preprint: 1907.11692, 2019. http://arxiv.org/abs/1907.11692.