Hans Journal of Data Mining 数据挖掘, 2021, 11(3), 135-149
Published Online July 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.113013

Hans汉斯

一种基于FinBERT-CRF命名实体识别模型的证券领域知识图谱构建框架

任秋宇，刘佳芮

同济大学，上海
Email: qiuyu.ren@tongji.edu.cn, rae1995liu@outlook.com

收稿日期：2021年4月18日；录用日期：2021年5月18日；发布日期：2021年5月27日

摘　要

随着信息媒介的转变以及人们对金融领域逐步的关注，证券领域新闻资讯信息的传递频率达到了前所未有的水平，而在当今金融领域缺乏一种能够可视化展示证券领域企业实体之间情感影响关系的建模方法。针对该问题，本文首先提出了一套实时的定向爬虫框架来获取所需的证券领域新闻文本，其次针对新闻文本设计了一种基于FinBERT-CRF的命名实体识别模型，最后结合市场基本面提出了一种构建面向情感分类的证券领域知识图谱，为投资者以及投资机构提供了一定的参考价值。

关键词

命名实体识别，知识图谱

Knowledge Graph Construction Framework in the Securities Domain Based on FinBERT-CRF Named Entity Recognition Model

Qiuyu Ren, Jiarui Liu

Tongji University, Shanghai
Email: qiuyu.ren@tongji.edu.cn, rae1995liu@outlook.com

Received: Apr. 18th, 2021; accepted: May 18th, 2021; published: May 27th, 2021

Abstract

The frequency of news and information transmission in the securities has reached an unprece-

文章引用: 任秋宇, 刘佳芮. 一种基于 FinBERT-CRF 命名实体识别模型的证券领域知识图谱构建框架[J]. 数据挖掘, 2021, 11(3): 135-149. DOI: 10.12677/hjdm.2021.113013

dented level with people's gradual attention to the financial domain and the transformation of information media. However, there is no modeling method that can visually display the emotional impact relationship between corporate entities in the securities field. In response to this problem, this article first proposed a real-time directional crawler framework to obtain the required securities field news text, then designed a FinBERT-CRF-based NER (named entity recognition) model for news. Finally, this paper proposes a knowledge map of securities field oriented to the sentiment classification based on the fundamentals of the market, which may provide a certain reference value for investors and investment institutions.

Keywords

Named Entity Recognition, Knowledge Graph

1. 引言

随着互联网便捷化的发展，以及新闻信息传播媒介的逐步转变，新闻资讯信息的传递速度达到了前所未有的水平。在金融证券领域的服务类别中，用户阅读金融类资讯的使用频率远超于其他金融服务，相对应，上市公司重大事件新闻的传播影响力也在快速增大。例如 18 年 7 月发生的长生生物疫苗事件，国家药品监管局早在同年 7 月 18 日便发布公告称长生生物疫苗研究中违反了相关规范，直到 7 月 22 日，一篇《疫苗之王》的公众号文章将长生生物推至风口浪尖，其股票价格一路下跌同时带动同概念生物疫苗相关股票价格一路狂跌。从类似事件的发生我们可以发现虽 Fama [1]于 1965 年提出了“有效市场假说”，但该假说的前提假定为：于证券市场参与者来说，人们均处于极其理性的情况下，同时掌握了金融证券市场完整的信息。但在真实的金融领域，由于中国的证券交易市场仍处于初期阶段且影响因素繁多，投资者往往难以自主对海量数据即时准确做出判断从而成为证券市场中的被动者。因此在计算机技术、深度学习逐渐兴起的背景之下，针对证券市场企业实体的情绪分类对于投资者以及投资机构来说尤为重要。近年来，在针对证券领域新闻文本的情感分类理论研究中，研究学者往往仅从文本化非结构性数据中获取相关影响因子或有效信息，而这一过程往往忽视了金融证券市场是一个综合整体的实体存在，对当前证券领域缺乏准确普遍的建模方法。故在本文中我们通过搭建面向情感分类的证券领域知识图谱，使证券领域实体之间的关系图像化，即为通过构建图边关系进一步对市场企业实体的情绪进行有效的分析打下基础。

在对金融证券领域的文本信息转换为知识图谱搭建中，首先需要考虑的问题是如何获得证券领域内的新闻文本和基本面数据，以及如何实现将其文本表述转换为计算机能够识别的表示方法。在传统的机器学习方法中，常需要充足的数据量来支撑参数调整优化，然而在特定专业领域，模型学习的效率往往由于上述假设过于严格而难以达到很好的效果。事实上，这些领域可用数据量缺失，使得训练样本往往不足以供复杂的机器学习模型进行训练从而得到一个可靠的生成预测模型。因此在本章，我们首先构建实时的定向爬虫框架来获取所需的证券领域新闻文本以及半结构化企业基本面数据。

其次本文针对非结构化数据证券领域新闻文本，提出基于命名实体识别的方法对证券领域知识图谱进一步情感分类标注的扩充。虽然命名实体识别模型搭建在现有的解决方案中已较为普遍，但由于证券领域新闻文本相关数据常存在一系列专有名词，大多数仅通过人为设立规则和模板方法为主，统计学习方法多作为辅助决策来完成。针对证券领域有标签数据常存在人工标注慢等此类问题，在本文中我们提

出了一套针对新闻文本的基于迁移学习的命名实体识别模型的搭建框架，通过同领域不同任务的参数迁移方法有效解决了该类问题。

最后，针对现有的证券领域中缺乏一种较为准确普遍的建模方法去体现企业实体之间的情绪表现关联，我们基于已爬取的半结构企业市场基本面数据，提出了一套证券领域知识图谱的构建框架。另外考虑到金融证券领域的特殊性，股市参与者对实体企业的信心往往能迅速表现于相应股价表现行为上。故在文中，我们将证券领域知识图谱与新闻中基于实体识别所得实体的市场涨跌表现进行融合，生成了面向情感分类的证券领域知识图谱。

2. 金融财经新闻文本数据的收集

在本章中，我们主要提出了一种针对证券领域的定向实体爬虫框架，为后续研究做出数据支持。之后对数据的分布与基本类型进行分析，设定规则为后续模型制作样本提供数据基础。

2.1. 基于上市企业数据的获取

随着网络信息的快速膨胀，如何准确快速的提取互联网信息已经成为一大新型挑战，爬虫技术即人为构造一套框架可以自动的从互联网中定向获取所需信息。在定向获取信息的过程中，人们往往会定义一套规则选择性下载目标信息，即常讲的定向爬虫。

如图 1 所示即为我们本节的金融领域非结构化新闻文本爬虫框架。首先，我们需要确定数据信息来源，并分析所需数据结构。本文中我们不仅仅要获得新闻文本等非结构化数据并存储至本地，同时需要获得其实体在股票中的基本面信息，例如地点、概念、板块等作为后续知识图谱构建关系实体的基础。根据爬虫的难易程度、新闻发布方的权威性等方面，我们在新闻资讯类非结构化数据方选取了以新闻权威性为主要竞争力的东方财富发布的 Choice [2]金融终端，作为本文的信息来源方。至于基本面数据半结构化数据，存储模块中我们设置了接入 Tushare [3]的 API 端口函数存储至本地。

Figure 1. A crawler framework for unstructured news text in financial domain
图 1. 金融领域非结构化新闻文本爬虫框架

其次，在我们从互联网直接获取非结构化数据的过程中，我们通过 Fiddle 记录下客户端与服务器交互过程中所有的 HTTP 请求，并对网络请求(request)和回复(reply)进行进一步的分析，查找得到了我们所需网页的请求包，我们相应进行伪装处理模拟器 header 进一步的访问，并依据预定义的爬虫规则从互联网上下载指定的数据。此外在制定获取网页的函数时，我们同时要考虑并行、执行频率等相关因素。由于本文涉及数据的并行化抓取，采用 Python 多线程模块处理所需定向 URL。在解析网页源代码中，网页文件一般以代码段的形式呈现，在经过 Service 处理后通过设计 URL 循环函数对定向更改参数的每一条链接进行数据抓取。利用 DOM 解析器，依据预定义的规则如取<p>标签下的可用文本将信息自动化地抽取出来。由于网页的结构复杂多样，基于规则的自动化抽取往往会产生一些错误，因此还需要清洗模块修正解析的结果。如图 2 所示，其网站的 HTML 中并没有明显区分标题与正文的不同，但文中有“【xxx】”的标记如“【中标项目】”来区分不同的文本板块。

最后需要对我们所获得的数据进行持久化的存储并清洗。在本节中我们主要存储与 mysql 数据库中，运用 python 中 pandas.tosql 进行批量存储。主要保存为字典类别的方式方便后续数据处理中转换为 dataframe 进行进一步操作。当我们存储到数据后，进一步清洗数据于 python 中的 scrapy 模块 ItemLoader

中进行，定义一个默认的全局输出处理器，再通过 TakeFirst 函数取出相应文本段，若为空，便丢弃整篇新闻文章。

Figure 2. Correspondence diagram of HTML page structure and key text
图 2. HTML 网页结构与关键文本对应图

2.2. 数据的分类与分析

在对本文的数据进行分类分析时，需要分别对半结构化数据以及非结构化数据两个场景进行考虑。

2.2.1. 半结构化数据的处理

首先我们对企业实体在股票领域的基本面半结构化数据进行预处理。通常的证券类数据包含日间交易数据、所在地域、所属行业、主营业务等基本面数据。在考虑相关方面数据时，主要目的为进一步发现证券领域企业实体间的隐藏信息。例如，当某日的白酒行业爆出“黑马”时，相应白酒类股票实体均表现优良。故在本文第三章构建令全领域知识图谱时，我们将行业、主营业务等概念同样以“实体”加入至知识图谱中。具体数据存储格式如表 1 所示：

Table 1. Example of crawler data storage format
表 1. 爬虫数据存储形式示例

ts_code	name	industry	chairman	province	city	employees
000001.SZ	平安银行	银行	谢永林	广东	深圳市	34204.0
000002.SZ	万科 A	全国地产	郁亮	广东	深圳市	133455.0
000004.SZ	国农科技	互联网	黄翔	广东	深圳市	251.0
000005.SZ	世纪星源	环境保护	丁芃	广东	深圳市	680.0
000006.SZ	深振业 A	区域地产	赵宏伟	广东	深圳市	385.0

具体为知识图谱建模所作的数据映射转换见第 3.1 节。

2.2.2. 非结构化数据的处理

证券类新闻是了解该时刻或当日重要事件的主要通道，在证券领域又称为另类宏观数据，它包含了企业、上市公司实体或相关概念的当日事件等信息，其中含有大量的金融证券领域的实体词语。在本节中我们主要对该类数据进行进一步的分类标签化预处理，为后续的 FinBERT-CRF 命名实体识别模型框架提供数据基础。具体映射前后区别如表 2 所示。

Table 2. Example of data format after mapping operation
表 2. 映射后数据样式示例 1

《北京大宗商品交易所(即北商所)诈骗百姓数十亿资金！》									
	标签		标签		标签		标签		标签
《	O	易	M-ORG	所	O	数	O	》	O
北	B-ORG	所	E-ORG	)	O	十	O		
京	M-ORG	(	O	诈	O	亿	O		
大	M-ORG	即	O	骗	O	资	O		
宗	M-ORG	北	O	百	O	金	O		
交	M-ORG	商	O	姓	O	！	O		

在本章模型的输入中，为了进一步强调输出类别标签之间的依赖性，我们将输入文本样本做出了以下几点约束：

- 实体标记的起始字符需为“B_”；
- 实体标记的中间字符需为“M_”；
- 实体标记的结束字记需为“E_”；
- 同一实体标记中间不能出现不同的类别标记，如“B_LOC, M_ORG”的标记连续出现是不被允许的；
- 其余情况均被标记为“O_”。

具体新闻文本数据的预处理(news preprocessing)的伪代码算法如下所示(算法 1)。

Algorithm 1. News text preprocessing algorithm
算法 1. 新闻文本预处理算法

Algorithm: News preprocessing
//输入：原始语料、规则限制以及标的词典
Input: raw text, limit, tag dict.
//输出：预处理后的带标签词语列表
Output: a list of marked words

// 把用户输入的字典与系统字典合并
1： *dictionary* ←combine *customized dictionary* with *universal dictionary*
// 遍历词语列表，删除无效单词
2： **for***word* **in** *word list* **do**
3： **if***word* is a stop word **then** delete it
// 根据模型输入的限制，把长的新闻词语列表截断，短的零或补全
4： **if** length of *word list*>*limit* **then** *word list* ← *word list*[0 : *limit*]
5： **else if** length of *word list*<*limit* **then** fill *word list* with “” until its length equals *limit*
// 遍历训练样本词语列表，根据标片标注训练样本的标识
6： **if***train sample word list == tag word list* **then** *word list add - B-ORG/M-ORG/E-ORG*
7： **return** *word list*

3. 基于 FinBERT-CRF 框架的证券领域新闻实体识别模型

为了进一步扩充证券领域知识图谱的情感分类属性，本章针对非结构化证券领域新闻文本，构建了一种基于 FinBERT-CRF 的证券领域新闻文本命名实体识别模型，用于识别新闻中的专有企业实体，以结合市场当时情绪扩充知识图谱中实体的情感表现。由于所爬取的证券领域新闻文本未自身带有企业实体标注，故我们采用了 DataFountain 金融领域实体新闻任务数据集进一步训练本节所提出的模型，并通过已训练模型对证券领域新闻文本进行实体识别，而 DataFountain 数据集中带有企业实体标注的规模较小，

故在模型 BERT 部分的初始参数设置时，通过复用了 FinBert [4]中在金融领域情感分析任务中已训练参数，有效减少数据规模小所带来训练不佳等影响，同时在参数优化阶段我们通过 CRF 损失函数对模型进行训练。本节将首先就本章模型 FinBERT-CRF 模型框架展开介绍，之后通过对比实验训练该命名实体识别模型，最后对其算法进行进一步分析并讨论结果证明模型算法的有效性。

3.1. 模型结构

在基于 FinBERT-CRF 的金融领域中文命名实体识别模型，我们首先设计 BERT 框架作为编码器并生成词向量，之后通过 Linear 层进行非线性映射生成概率矩阵，最后通过 CRF (Conditional Random Field) 损失函数[5]对输入标签进行学习反馈得到最终结果。模型框架如图 3 所示。

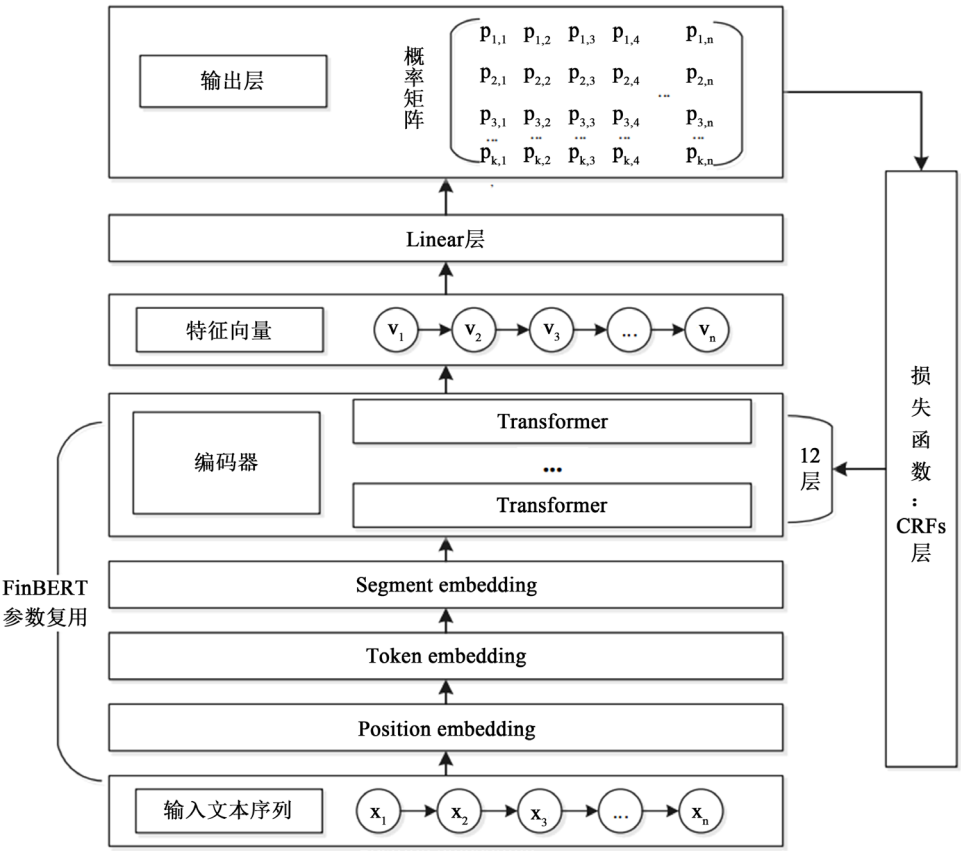

Figure 3. Model structure of FinBERT-CRF
图 3. FinBERT-CRF 模型框架

从图 3 可以看出，模型共分为三个模块，分别为输入层、编码层以及输出层。在输入层中我们将输入的中文文本样本，转换为模型可识别的文本格式输入文本序列$(x_0, x_1, x_2, \cdots, x_N)$。在编码层中我们主要将输入文本序列转换为已处理的特征向量格式。输入层中的文本$x=(x_0, x_1, x_2, \cdots, x_N)$以单个汉字与相应标签如“司 O-ORG”的格式作为一个单位的n个单位输入序列。在本节的编码层中我们主要使用了 BERT 模型思想，且在 BERT 模块的参数设置中，我们复用了 FinBERT 训练所得参数。FinBERT [4]模型为一种基于 BERT 的金融领域文本模型，在金融情绪分类的任务中的 Financial Phrasebank 数据集达到了 86%的准确率。BERT 在文本特征向量生成中相较于仅考虑单向训练的语言模型，其同时训练了每个单词上下文。其中编码层主要包含了两个部分，分别为嵌入层与 Transformers [6]所组成的编码器层。

在嵌入层中除了输入层文本之外，存在段嵌入层(Segment embedding)、符号嵌入层(Token embedding)以及位置嵌入层(Position embedding)。段嵌入层的目的是使模型对文本中不同的句子关系进行区分，在生成过程中嵌入 A、B 符号表示区分多个句子在整篇文章中的奇偶性。符号嵌入层的目的是在针对每个句子的处理中，从词典中查询对应单词的词语向量，并该嵌入层在每个样本前端插入[CLS]标识，同时在每句话的最后加入[SEP]标识来表明每个句子的结束边界，这样句子的语义特征均可通过 CLS 训练表达。位置嵌入层的目的为识别文本中单词的前后依赖关系，我们在最后一层嵌入了位置向量层去传递模型相应单词的位置信息。在三层嵌入层之后我们将输入向量进入标准编码层，本文中我们通过使用 12 个 Transformer 层叠加聚合并同时输出生成了了一组特征向量 $v=(v_0,v_1,v_2,\cdots,v_N)$。

在输出层中我们主要将已生成的特征向量将其转换为对应标注的概率矩阵并经过损失函数进一步调整训练相关参数。在特征向量生成后进入 Linear 层，将编码器部分输出的特征向量做映射变换生成不同的 logits 向量，随即进入 softmax 层转换为概率值，最后便得到文本输出的各个标签对应的概率矩阵即为我们的输出结果。在输出层得到概率矩阵后在本节我们通过设计了 CRFs 算法作为损失计算函数，根据 CRFs 计算出的条件概率得到了最大化对数似然函数并使用梯度下降法，根据损失函数反向学习模型参数输入至我们的编码器中，优化相应调整有关参数。在训练结束后在不同的文本位置处输出结果选择其中概率最高的标签类别作为当前的预测结果并进行解码输出即为我们的标签结果。

3.2. CRF 损失函数

在本文中我们通过 CRF 来构建模型的损失函数，CRF 损失函数共由两个部分组成，真实标签所得的概率分数以及所有标签的总分数。假设我们的数据集中有以下的类别标如表 3 所示。

Table 3. Data set category table
表 3. 数据集类别表

标签	序号
B-ORG	0
M-ORG	1
E-ORG	2
O	3
START	4
END	5

若存在一个包含六个单词的句子，可能的类别序列为存在 46656(6^6)种可能性，假设每种路径的可能概率为 P_i，则路径总分如式(1)所示。

$$P_{all}=P_1+P_2+\cdots+P_{46656}=e^{S_1}+e^{S_2}+\cdots+e^{S_{46656}} \tag{1}$$

若其中第 m 个标签路径为真实标签路径，则 P_m 应为所有概率序列中最高的概率值，定义的损失函数如式(2)所示。

$$Loss=\frac{P_m}{P_1+P_2+\cdots+P_{46656}} \tag{2}$$

其中由于 $P_m=e^{S_m}$，则在计算真实概率值中，e 为固定常量参数则需计算 S_m。S_m 主要是由两个部分所构成，分别为发射分数与转移分数，该分数即有 CRF 层中的状态转移矩阵得以计算。

3.3. 实验结果

在本章中我们主要介绍了本章的实验环境，同时对对比实验研究中涉及的训练数据集进行了进一步的介绍，最后展示了我们的对比实验结果，并表明本文所提出的模型框架以及实验思想在小规模证券领域命名实体识别的任务中得到了优异的表现。

3.4. 实验数据

本章中主要涉及到两个训练数据集，分别为通用细粒度命名实体识别数据集 Cner、数据科学竞赛平台 DataFountain 金融类文本新实体识别数据集。具体说明如下：

1) 通用细粒度命名实体识别数据集 CNER [7]

Cner 数据集为已处理的中文通用数据集，共包含十种不同的类别，包括：人名(name)、组织(organization)、地址(address)、公司(company)、政府(government)等，每组样本以输入的原始单个文本和标记为键值对的组成序列。其中训练集包含 3820 个样本，验证集包含 462 个样本。具体如下表 4 所示：

Table 4. CNER Data sample
表 4. CNER 数据示例

	标签		标签		标签		标签		标签		标签
吴	B-NAME	，	O	级	M-TITLE	院	O	邮	B-ORG	研	B-TITLE
重	M-NAME	大	B-EDU	高	M-TITLE	特	O	电	M-ORG	究	M-TITLE
阳	E-NAME	学	M-EDU	工	E-TITLE	殊	O	部	M-ORG	所	M-TITLE
，	O	本	M-EDU	，	O	津	O	侯	M-ORG	副	M-TITLE
中	B-CONT	科	E-EDU	享	O	贴	O	马	M-ORG	所	M-TITLE
国	M-CONT	，	O	受	O	，	O	电	M-ORG	长	E-TITLE
国	M-CONT	教	B-TITLE	国	O	历	O	缆	M-ORG	。	O
籍	E-CONT	授	M-TITLE	务	O	任	O	厂	E-ORG		

2) 数据科学竞赛平台 DataFountain 金融类文本新实体识别数据集[8]

该数据集原始格式以 csv 文件的格式提供，为金融类网络文本包括金融类新闻，每条数据包括标识号码(id)、新闻标题(title)、新闻内容(text)、未知实体列表(unknownEntities)即实体列表五种文本，且提供的比赛数据中训练集与测试集分别均含有 10,000 条中文新闻文本。由于该类数据集已标记了文本相关实体，故我们对该原始数据集做统一的数据分类与分析处理同 3.1 节框架方法作半结构化数据预处理，得到我们的样本总集并将其按照 7:2:1 的方式分为训练集、验证集与测试集。

3.5. 评测指标

本章实验的评价系统采用命名实体识别任务中常用的三类指标：精确率(P, Precision)、召回率(R, Recall)和 F1-Measure ($F1$)，具体公式分别如下式(3)、式(4)、式(5)所示。输出的预测标注的结果为 $S=\{s_1,s_2,s_3,\cdots,s_n\}$，人工标记的结果为 $T=\{t_1,t_2,t_3,\cdots,t_n\}$。在匹配的过程中，需要同时评价实体类别以及标记边界范围预测的效果，故本文中若该两个部分均预测准确，才作为正确预测计入评测系统中。

$$P=\frac{correct_{num}}{predict_{num}} \tag{3}$$

$$R=\frac{correct_{num}}{gold_{num}} \tag{4}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{5}$$

其中的 $gold_{num}$ 即表示我们人工标记的结果 T。

3.6. 实验结果分析

在构建本章基于 FinBERT-CRF 的证券领域命名实体识别模型框架中，部分原始模型参数选择如表 5 所示。

Table 5. Part of the model parameters
表 5. 部分模型参数

参数	参数值
隐层值(hidden_size)	768
隐层数目(num_hidden_layers)	12
注意力头数目(num_attention_heads)	12
隐层随机下降值(hidden_dropout_prob)	0.1
注意力层随机下降值(attention_probs_dropout_prob)	0.1
输入最大文本长度(max_position_embeddings)	512
标准化范围(initializer_range)	0.02
归一化范围(layer_norm_eps)	1e−12

在本章所提出的FinBERT-CRF证券领域实体识别模型框架中的BERT模型进行训练主要分为两个步骤，分别为预训练和微调(Fin-tune)。

在预训练的过程中，本文模型主要通过三种不同的方式：第一种，直接调取 Google 提出的 BERT 模型参数迁用至本章 BERT 模型中；第二种，我们通过在通用实体识别数据集即 Cner 数据集输入至 BERT 模型进行预训练，第三种，我们使用了 FinBERT 模型已训练参数加入至 BERT 模型之中即本文所提出模型。

在微调的过程中，使用上述四种情况得到的参数初始化 BERT 模型，后通过下游任务的目标基于 DataFountain 金融类文本新实体识别数据集对其进行小步幅的调整，并通过 CRFs 损失函数调优最终得到各类标签的概率矩阵。得到对比实验结果如下图 4 所示。

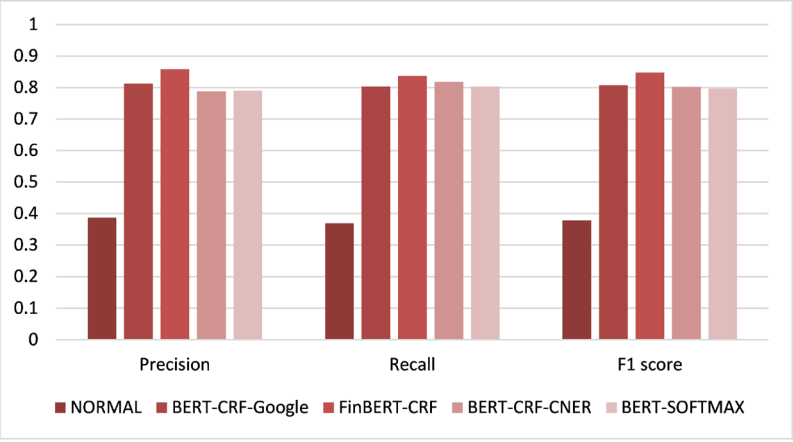

Figure 4. Comparison of experiments results
图 4. 实验结果对比图

其中如下表 6 所示，NORMAL 表示我们对 BERT [9]的语调参数作随机初始化，BERT-CRF-Google 表示我们直接应用 Google 提出的 BERT 模型参数预训练基础下使用小规模数据集 Kaggle 比赛中的文本在 BERT-CRF 模型架构上进行训练得到的结果；FinBERT-CRF 表示我们应用 FinBERT 模型参数并再 DataFountain 数据集训练学习得到的结果；BERT-CRF-CNER 为上述所提在通用数据集预训练 CNER 基础得到的结果；BERT-SOFTMAX 为不通过 CRF 作为损失函数的 BERT-Linear 模型。

Table 6. Experimental results on different methods
表 6. 实验对比结果

单位：%	Precision	Recall	F1 score
NORMAL	38.72	36.94	37.81
BERT-CRF-Google	81.18	80.31	80.74
FinBERT-CRF	**85.78**	**83.65**	**84.70**
BERT-CRF-CNER	78.77	81.77	80.24
BERT-SOFTMAX	78.97	80.31	79.63

从上述实验可以看出，FinBERT-CRF 相较 BERT-SOFTMAX 模型可以提升 6.81%的准确率，故本节中提出的以 CRF 为损失函数可以有效的提升模型的训练效果。在对比 NORMAL 与本文所提基于金融领域新闻预训练的 FinBERT-CRF 的过程中，我们可以发现同领域不同应用任务模型的参数可以有效解决特定领域数据不足的问题。在对比 BERT-CRF-Google、BERT-CRF-CNER 与 FinBERT-CRF 的过程中可以发现同领域迁移学习模型较跨领域迁移学习表现更为优异。

从上述实验中我们可以发现，本章所提出的基于 FinBERT-CRF 框架的证券领域命名实体识别模型可以在迁移学习的思想下可以有效并优异的完成小规模样本下证券领域命名实体识别任务。

4. 基于 FinBERT-CRF 命名实体识别的证券领域知识图谱的构建

在前两节的数据准备以及 FinBERT-CRF 命名实体识别模型的基础上，本节主要介绍本章基于金融证券领域的知识图谱的设计以及基于 FinBERT-CRF 对证券领域知识图谱情感分类的扩充算法。

4.1. 初始证券领域知识图谱的设计

本小节主要通过已爬取的企业于股票市场中的基本面数据构建初始的证券领域知识图谱。三元组作为知识图谱的基本结构，在进行本文领域知识图谱的设计的过程中是不可忽视的一部份。根据本文已爬取数据的实际情况，主要设计了两种相应的存储结构。第一种为“实体－属性－具体属性”，该类三元组具体表现为“一对一”的形式即每一个实体的特定属性仅有唯一对应的值存在。例如实体姓名(name)、是否退市(list_state)、总资产(reg_capital)、省份(province)、城市(city)、主营业务(main_business)为唯一属性映射形式。第二种为“实体－关系－实体”，该类三元组表现为“一对多”的形式即通过关系映射，某一实体的某一特定关系存在多个实体。如下表所示，表 7 列举了某一企业实体的实体属性“一对一”映射关系。

本文在构建证券领域知识图谱中创建了 code (股票代码)、name (企业姓名)、list_state (上市情况)、经营规模(reg_capital)、管理者(chairman)、province (省份)、city (城市)七个节点属性标签，位于(located)、属于(belong)、主营(main_business 主营业务)三个关系属性标签。如表 7 中 list_state 中的 L 表示该企业实体正常上市中，D 表示该企业实体已经退市。三个实体属性和省份、城市、行业、概念、主营业务五个实体之间的关系。

Table 7. Examples of “one-to-one” mapping
表 7. 属性“一对一”映射关系举例

实体	属性	属性值
000024	Name	招商地产
000024	list_state	D
000024	reg_capital	257595.0754
000024	province	广东
000024	city	深圳
000024	main_business	房地产开发经营
000024	chairman	孙承铭

4.2. 证券领域知识图谱的构建

本节为进一步实现“实体－属性－属性值”和“实体－关系－实体”的结构特性，需对抓取到的数据进行预处理，为构建初始证券领域知识图谱提供数据基础。首先将 industry 表中的一对多的映射信息转换为“一对一”映射的数据表中。例如 industry 表之间的转换，我们需要将表 8 通过 Pandas 与 math 包等处理为表 9 中“一对一”映射方式。

Table 8. Example of “one-to-many” mapping table for industry
表 8. Industry 表“一对多”映射表举例

Id	股票代码	行业	股票代码
14863	300757	专用机械	300724, 300776, 300812, 601226, 603036
14864	300681	汽车配件	601127, 601799, 603013, 603035, 603085
14865	002649	软件服务	300624, 300634, 300645, 600289, 600476

Table 9. Example of “one-to-one” mapping table for industry
表 9. Industry 表转换为“一对一”举例

Id	实体	关系	实体
43508	300757	主营	专用机械
43509	300724	主营	专用机械
43510	300776	主营	专用机械
43511	300812	主营	专用机械
43512	601226	主营	专用机械
87501	601127	主营	汽车配件
87502	601799	主营	汽车配件

通过该种方式，本节共创建了股票(code)、城市(city)、省份(province)、主营业务(main_business)、概念(notion)五种节点，股票姓名－城市(name-city)、城市－省份(city-province)、股票姓名－主营业务(name-main_business)、股票姓名－所属概念(name-notion)、股票姓名－省份(name-province)五种实体关系映射表。最后将预处理后的表格数据导入 neo4j [10]图数据库中。Neo4j 数据库提供了 Cypher 语句进行逐条导入，该种方式可以实时插入数据且修改数据较为灵活，同时也可以通过 LOAD CSV 语句进行基础的

csv 文件多批次插入。本文中我们在 Centos7 系统下将处理过的节点、关系文件通过 neo4j-import 的方法进行批次插入，得到的部分可视化呈现如图 5 所示。

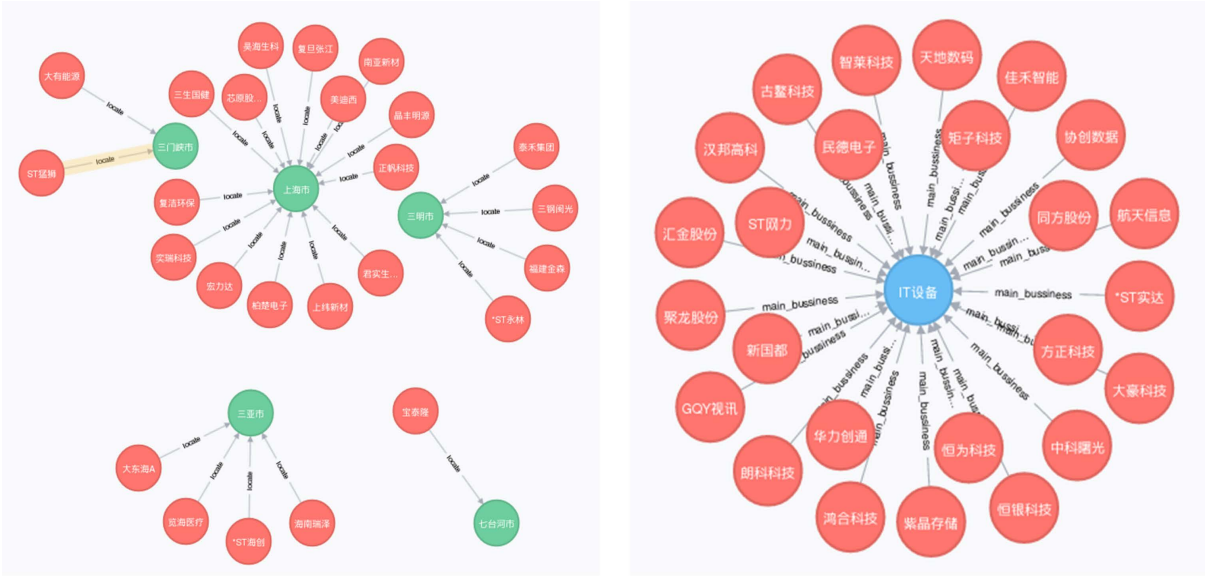

Figure 5. Visulization of knowledge graph in the securities field
图 5. 证券领域知识图谱可视化部分呈现

在本节所构建的初始证券领域知识图谱中，共包含了 4673 个实体节点、18,986 个属性描述以及 13,340 条边关系。

5. 基于 FinBERT-CRF 的证券领域知识图谱的构建

为了进一步实现对于企业实体节点的情感分类任务，本节主要介绍了在已构建的初始证券领域知识图谱基础之上，基于 FinBERT-CRF 构建面向情感分类的证券领域知识图谱的构建原理以及算法实现。

在构建了初始证券领域的知识图谱之后，本节将主要就面向情感分类的证券领域知识图谱的构建的流程展开介绍。对于非结构化新闻文本数据，我们基于 FinBERT-CRF 命名实体识别模型，将爬取的相关新闻文本作为训练样本送入体模型中，即可训练得到对应文本相关企业实体，之后通过如图 3 所示流程为模型所输出的实体打上情感标签，生成面向情感分类的新闻实体数据集，最终与知识图谱融合生成最终的面向情感分类的证券领域知识图谱，具体流程如图 3 所示。

如图 6 所示，我们首先将爬取的新闻文本数据送入基于已训练的 FinBERT-CRF 命名实体识别模型，分别将其中相关企业实体进行识别并标注，得到新闻样本中的相关实体集 $\{A, B, \cdots, M, \cdots\}$。

其次我们基于新闻发布时间点以及相关企业实体，通过实际的股票市场中新闻发生第二日后具体涨跌表现，将其标注为相应新闻实体的情感表现。在本节中我们考虑到新闻事件的时效性，选择了一日后股票的表现作为其情感表现的评估标准。

例如在 2020 年 4 月 16 日晚间公布的一条新闻中“省广集团：暂未开展 RCS 相关业务”，则其中得到的企业实体为“省广集团”，在证券市场中“省广集团”的于 2020 年 4 月 17 日的涨跌幅为−8.16%，故该条发生于 4 月 16 日的新闻我们将其打为负情绪。以此类推，在本数据集中我们假定下述三种情况为我们制作标签的规则：

- 若第二日的涨跌幅大于 1.5%，则为正类情感标签；
- 若第二日的涨跌幅小于−1.5%，则为负类情感标签；
- 若第二日的涨跌幅位于−1.5%至 1.5%之间，则为中性情感标签。

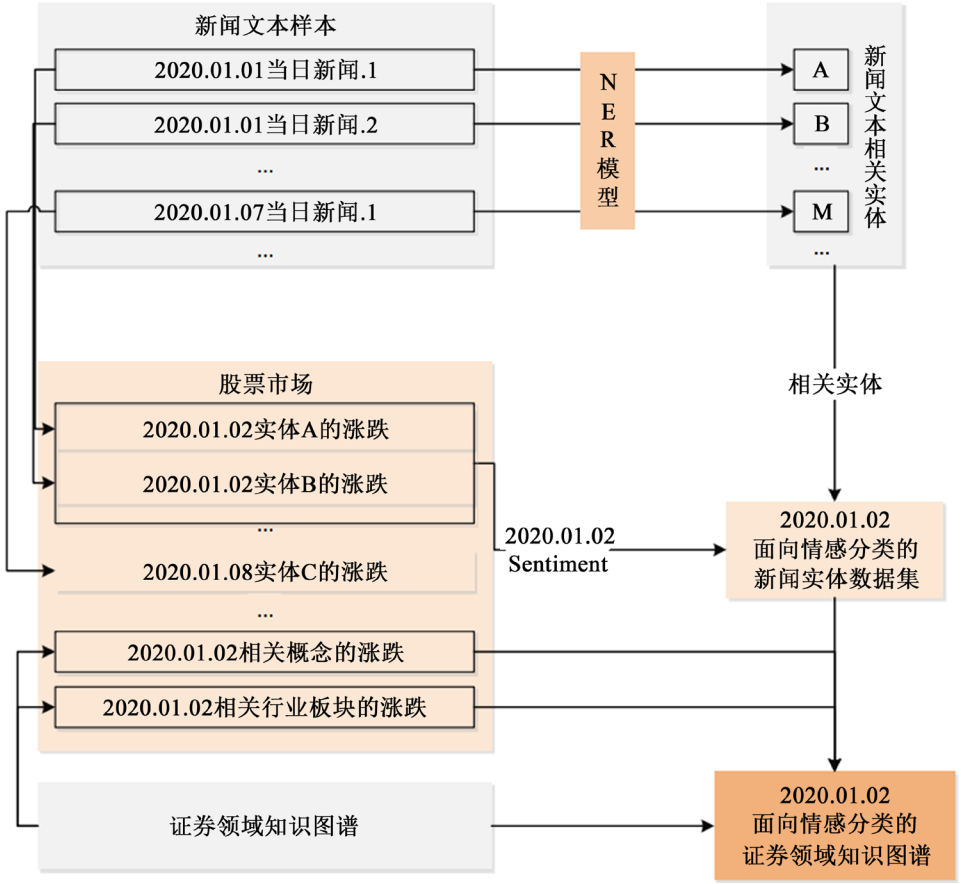

Figure 6. Construction process of Knowledge graph FinKG-EMO in Securities field
图 6. 面向情感分类的证券领域知识图谱 FinKG-EMO 构建过程图

之后按照该类情感分类的标签方法对其进行迭代即可得到对应新闻文本中实体的当日市场相关情绪值，同时我们将证券领域知识图谱中的相关概念以及板块如“新能源”“白酒”等节点在第二日的市场表现仍遵从以上规则进行情感分类标签处理。

最后结合已构建的知识图谱三元组生成在时间区间内每一日的证券领域知识图谱相即生成一系列的相关样本结构图数据集，同时考虑到新闻样本的数据集规模较小，我们通过将一组十张连续日的知识图谱将节点情绪进行了如式所示的拼接。

$$E_N = \sum_{i=0}^{N} \alpha_i E_i / N \tag{6}$$

如式(6)所示，其中 E_i 代表第 i 日企业节点E在市场中的情感分类，α_i 代表第 i 日企业节点 E 对第 N 日节点E的情绪权重，我们在本节中我们设置 $N \le 10$ 且 α_i 中 i 越小相应的权重越小，最终生成了基于 FinBERT-CRF 模型的面向情感分类的证券领域知识图谱数据集 FinKG-EMO，具体实现算法如算法 2，且相应生成的具体实体节点情感三元组样本示例如表 10 所示。

Algorithm 2. FinBERT-CRF-based knowledge graph construction algorithm
算法 2. 基于 FinBERT-CRF 的证券领域知识图谱构建算法

Algorithm: Knowledge Graph Based on Finbert-CRF Construction Algorithm
//输入：金融领域新闻，已训练的 NER 模型，知识图谱，初始化第 i 日情感权重
Input: financial news list, NER trained modeling, knowledge graph G=(V,E), i=0
//输出：面向情感分类的知识图谱
Output: G with emotion

// 将金融领域新闻送进已训练的 NER 模型中得到实体集
1： *entities list* ←train *financial news list* with ***NER trained modeling***
// 遍历新闻列表，将新闻日期与已训练的实体进行合并
2：for*news***in***news list* **do**
3： *entity with time* ←combine *news date* with *corresponding entity*
4：append*entity with time***to***entities list*
5: end for
// 根据实体与日期的市场表现，为相应实体标注市场情绪表现
6：for*entity, time***in***entities list* **do**
7： **if** *entity performance in security market >2%* **then** mark as *1 (positive)*;
8： **if** *entity performance in security market <-2%***then** mark as *0 (neutral)*;
9： **else** mark as *-1 (negative)*;
10: end for
// 根据日期遍历初始搭建的知识图谱中的实体并标识相应的情感分类并将连续 10 日内的知识图谱归为一个集合
11：for *entity, time***in***entities list* **do**
12： **if** *entity* **in** *G=(V,E)* **then** mark *node* with time, 0/1/-1;
14： **if** *time* **in** *10 days* **then**append *G with mark* to G_{10}
14: end for
// 将 10 日内的知识图谱集合合并为一个知识图谱
15：for*G***in***G_{10}* **do**
16: t = 0
16:for *entity, mark***in***G***do**
17: **If** t < 10**do**
18:i = i + 0.1, t = t + 1
19： *sentiment of entity* =i* $sentiment_i$+ *sentiment of entity*
20: **end for**
21: **append** *sentiment of entity*/10 **to** *G*
22: **end for**
23: return*G with sentiment list*

Table 10. An example of storage representation of entity nodes of knowledge graph in the securities field
表 10. 面向情感分类的证券领域知识图谱实体节点存储表示举例

Id	Name	Emotion
000089	深圳机场	0
000090	天健集团	−0.6
000096	广聚能源	−0.8
000099	中信海值	−1
000100	TCL 科技	0
000150	宜华健康	0.3

除此之外，为了进一步充实面向情感分类的证券领域知识图谱中节点的特征信息，我们将已搭建的知识图谱 FinKG-EMO 中的企业节点基于上述 FinBERT 模型编码嵌入层部分生成了一系列的情感特征向量附于知识图谱并进行初步拼接，为之后的模型训练提供了有效的数据信息。

6. 小结

本文针对当前证券领域缺乏准确普遍面向关系的可视化建模方法在一定假设下提供了可能解决方案。

针对获取数据方面，本章提出了一套完整的金融财经新闻文本数据的爬虫框架，可基于该框架及时准确地从互联网中获取大量的上市企业新闻文本数据以及非结构化企业基本面数据，为进一步的命名实体与知识图谱的搭建提供数据支持。对于非结构化新闻文本，本节在特定领域样本规模小的假设前提下提出了基于 FinBERT-CRF 框架的证券领域新闻命名实体识别的模型，通过复用同领域情感分析 FinBERT 模型中的参数，实现从小规模证券领域文本中提取新闻相关实体并通过对比实验证明该模型框架的有效性。最后，基于上述爬虫所得半结构化数据所构建的初始证券领域知识图谱以及基于 FinBERT-CRF 命名实体模型所得相关实体的市场情感表现搭建了面向情感分类的证券领域知识图谱，为本文下述基于图的情感分类模型研究提供了支持。

综上所述，本文主要解决了如何能够及时准确地从互联网中获得所需的证券领域新闻文本的问题、如何能在小规模新闻文本数据集上准确高效地识别证券新闻涉及的相关企业实体的问题以及如何结合实际应用场景根据已有的半结构化数据和相关实体的市场表现搭建有效的领域知识图谱的问题，分别提出了一套基于上市企业的定向爬虫框架、基于 FinBERT-CRF 的证券领域命名实体识别模型以及基于命名实体识别模型的一套构建面向情感的证券领域知识图谱框架，为后续基于图结构的情感分类研究展开提供了基础，并为类似特定领域的情感或文本类研究提供了实践参考。

参考文献

[1] Malkiel, B.G. and Fama, E.F. (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, **25**, 383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

[2] http://choice.eastmoney.com/

[3] https://tushare.pro/

[4] Araci, D. (2019) FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models. arXiv:1908.10063 [cs.CL]

[5] Panchendrarajan, R. and Amaresan, A. (2018) Bidirectional LSTM-CRF for Named Entity Recognition. *The 32nd Pacific Asia Conference on Language, Information and Computation* (*PACLIC* 32), 2019.

[6] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[7] Wu, F., Liu, J., Wu, C., *et al.* (2019) Neural Chinese Named Entity Recognition via CNN-LSTM-CRF and Joint Training with Word Segmentation. *The World Wide Web Conference*, San Francisco, May 2019, 3342-3348. https://doi.org/10.1145/3308558.3313743

[8] https://www.datafountain.cn/competitions/361

[9] Devlin, J., Chang, M.W., Lee, K., *et al.* (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[10] López, F.M.S. and De La Cruz, E.G.S. (2015) Literature Review about Neo4j Graph Database as a Feasible Alternative for Replacing RDBMS. *Industrial Data*, **18**, 135-139. https://doi.org/10.15381/idata.v18i2.12106