第 29 卷 第 1 期 上海大学学报（自然科学版） Vol. 29 No. 1
2023 年 2 月 JOURNAL OF SHANGHAI UNIVERSITY (NATURAL SCIENCE EDITION) Feb. 2023

DOI: 10.12066/j.issn.1007-2861.2308

基于 BERT 的金融文本情感分析模型

朱 鹤, 陆小锋, 薛 雷
(上海大学 通信与信息工程学院, 上海 200444)

摘要: 在金融领域, 越来越多的投资者选择在互联网平台上发表自己的见解. 这些评论文本作为舆情的载体, 可以充分反映投资者情绪, 影响投资决策和市场走势. 情感分析作为自然语言处理 (natural language processing, NLP) 中重要的分支, 为分析海量的金融文本情感类型提供了有效的研究手段. 由于特定领域文本的专业性和大标签数据集的不适用性, 金融文本的情感分析是对传统情感分析模型的巨大挑战, 传统模型在准确率与召回率上表现较差. 为了克服这些挑战, 针对金融文本的情感分析任务, 从词表示模型出发, 提出了基于金融领域的全词覆盖与特征增强的 BERT(bidirectional encoder representations from Transformers) 预处理模型.

关键词: 情感分析; 词嵌入向量; BERT; 词性特征; 命名实体识别
中图分类号: TP 391.1 **文献标志码:** A **文章编号:** 1007-2861(2023)01-0118-11

Emotional analysis model of financial text based on the BERT

ZHU He, LU Xiaofeng, XUE Lei
(School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: In the financial sector, more and more investors choose to express their opinions on the internet platform. These comment texts can fully reflect investor sentiment and influence their investment decisions and market trends. Emotion analysis as an important branch of natural language processing (NLP), which provides an effective research means for analyzing a large number of text emotional types in financial sector. However, due to the professional nature of domain-specific texts and the inapplicability of large label data sets, text emotion analysis in the financial field has brought great challenges to the traditional emotion analysis model. When the general emotion analysis model is applied to specific fields such as finance, its accuracy and recall rate are poor. In order to overcome these challenges, a BERT (bidirectional encoder representations from Transformers) preprocessing model based on full word coverage and feature enhancement in financial field was proposed for the emotional analysis task of financial text from the perspective of word representation model.

收稿日期: 2020-12-22
基金项目: 上海市科委基金资助项目 (19511105503)
通信作者: 薛 雷 (1963—), 男, 副教授, 博士, 研究方向为模式识别. E-mail: xuelei@shu.edu.cn

Key words: sentiment analysis; word embedded vector; BERT; bag-of-POS (part of speech); named entity recognition

在人工智能时代, 自然语言处理 (natural language processing, NLP) 技术引起了学术界和工业界的广泛关注, 而文本数据的情感分析是其重点研究方向之一, 具有很高的研究和应用价值. 金融领域中的股票、基金以及期货等中文文本数据与日俱增, 具有产生速度快、蕴含信息量大的特点. 如何能快速、准确地挖掘金融文本中的隐藏信息, 是目前迫切需要解决的问题. 因此, 训练出一个能够自动处理大量金融文本信息的模型, 在辅助投资者、金融投资机构进行投资参考以及政府掌握金融市场的舆论风向、分析投资者的态度等方面, 具有丰富的参考价值.

本工作针对金融文本中的情感分析任务, 从词表示模型角度进行创新, 提出了基于金融领域的全词覆盖与特征增强的 BERT(bidirectional encoder representations from Transformers) 预处理模型. 本工作具有以下两点贡献: ①针对金融领域的专业名词对原始 BERT 模型进行针对性的训练处理, 提出了一个金融领域全词覆盖 (whole word masking) 的 BERT 模型, 更加适用于金融文本情感分析任务; ②为了更好地分析与判断实例级的金融文本, 对任务本体进行细化, 进一步提取文本中的特征, 将词性标注特征、命名实体特征等多种金融特征与本工作提出的金融领域全词覆盖的 BERT 模型进行融合, 实现了更为精确的金融文本情感分析.

1　相关工作

面对海量的金融评论文本数据, 若仅仅依靠投资者进行阅读和分析提炼有价值的信息是不现实的. 一方面在时间上不可行. 金融市场极度复杂多变, 股值状态转移迅速, 依靠人工分析的方法具有严重的滞后性, 使得数据失去时间效力. 另一方面由于投资者的基础和背景不同, 每个人对市场变化的看法具有片面性, 个人不能对海量金融文本数据进行综合分析. 比如, 股民在进行选股时, 由于心理因素, 可能仅仅通过几个评论、帖子就轻易定下结论, 错失投资良机.

NLP 为快速分析大量金融评论文本的情感倾向提供了可能的解决方案. 情感分析是 NLP 的任务之一, 其定义是从文本数据中挖掘用户表达的观点和情感极性[1]. 在情感分析研究的早期, 大多数研究使用基于词典的方法[2-3], 根据预定义的词典类别将单词、短语或句子分类为组. 缺点在于其结果准确度受限于情感词典的质量和情感判断规则的搭配, 因此随着数据量的不断增加, 无法处理日渐复杂的文本情感分类问题.

已有研究证明, 基于机器学习的方法比基于词典的方法在情感分析任务上可以获得更高的准确率[4]. 一般来说, 基于机器学习的情感分析方法主要包括以下 3 个步骤: ①获取文本数据集; ②手动提取文本数据的情感特征; ③使用机器学习的算法进行训练分类. Hai 等[5]针对在线用户生成的评论和整体评价, 提出了一种新的概率监督的、面向情感的联合模型——有监督的联合主题建模方法 (supervised joint aspect and sentiment model, SJASM). 该概率模型建立在文档主题生成模型, 即隐含狄利克雷分布 (latent Dirichlet allocation, LDA) 模型基础上, 能够在统一的框架内兼顾方面层情感分析和整体情感分析. Singh 等[6]综合了朴素贝叶斯、J48 决策树、BF 决策树以及 oneR 算法这 4 种机器学习算法进行了文本情感分析. Al-Amrani 等[7]提出了融合支持向量机 (support vector machine, SVM) 与随机森林 (random forest, RF) 的 RFSVM 混合模型, 并将 RFSVM 混合模型在亚马逊商品评论数据集上进行了情感分类实验. 结果证明, 有监督学习的 RFSVM 混合模型结合了 RF 和 SVM 的优点, 相对

于其他算法具有更好的稳定性. Hai 等[5]还研究了股评篇章的结构, 并利用 SVM 进行训练, 提出了一种针对股评的情感分类方法. Singh 等[6]对金融新闻进行语义标注、提取特征, 然后分别使用朴素贝叶斯、SVM 和 k-近邻算法进行分类. 但是, 依靠手工提取特征的情感分析方法, 在面对复杂的文本特征以及大量的文本时, 仍存在一定的局限性.

基于深度学习的方法一般采用稠密、连续、低维度的向量表示文档和词语, 所以能解决数据稀疏问题[8]. 与此同时, 深度学习网络可以自动提取文本特征, 大大降低了文本特征构建的复杂性和不确定性. Man 等[9]探讨了深度学习方法在金融文本情感分析任务上的应用, 例如词嵌入、卷积神经网络 (convolutional neural network, CNN)、循环神经网络 (recurrent neural network, RNN)、长短期记忆 (long short term memory, LSTM) 网络和注意机制 (attention mechanism) 等, 并提出大型无监督语料预训练模型是未来可能的研究方向.

目前最新的情感识别里程碑是 Devlin 等[10]的 BERT 模型. 该模型已获得了 11 种 NLP 任务最先进 (state-of-the-art) 的结果, 其原因是在语言模型预训练阶段, 采用了基于自注意力机制的双向 Transformer[11]结构. 对比于语言模型嵌入 (embeddings from language models, ELMo)[12], BERT 模型的特征抽取器使用特征抽取能力更强的 Transformer 特征提取器代替 LSTM. 对比于生成式预训练 (generative pre-training, GPT)[13], BERT 模型将单向解码器更换为双向的 Transformer 结构, 解决了长文本上下文语义依赖的问题.

BERT 模型的预训练阶段包括两个任务: ①遮蔽语言模型 (masked language model, MLM), 根据自己特有的遮盖 (MASK) 语言模型预训练方式, 生成对每个输入单词 (token) 的上下文分布式表示; ②下句预测 (next sentence prediction, NSP) 模型, 预测金融语料库中的下一个句子. 在特定场景使用该预训练模型时, 不需要利用大量的语料来进行训练, 提高了时间效率, 具备较强的泛化能力, 对于特定的任务只需要添加一个输出层来进行微调 (fine-tuning) 即可. 例如, BERT 模型在无标记的数据集上进行转移学习, 可以在情感识别任务中获得更好的结果[10]. BERT 的变形结构 GPT-2, 借助庞大而多样的数据集和非常深的神经网络, 进一步显示了无监督预训练的强大功能, 可以达到情感分析任务的最高识别准确率[14].

2 模型设计

2.1 基于 BERT 的金融领域全词覆盖模型

在中文的 NLP 中, BERT 模型原有的分词方式会把一个完整的词组分解为若干个字. 例如, “上海大学” 一词在输入时会被拆分为 “上”、“海”、“大”、“学”4 个字. 在训练过程中, 这些字会随机被 [MASK] 替换. 显然, 这样的训练方式并不能使模型较好地学习到文本中的语义信息. 本工作采用金融领域全词覆盖的 BERT 预训练模型, 即当一个与金融领域相关词组中的部分字在训练过程中被 [MASK] 覆盖时, 则同属于该词组的其他字也会被相应的覆盖. 通过对这些词进行基于金融领域全词覆盖的 BERT 预训练后, 该预训练模型对于金融领域的任务将更为贴切、适应能力更强, 可以提取到更多的金融领域信息, 从而更好地解决金融领域评论中语义模糊、关键特征稀疏的问题.

为了使 BERT 模型更加适用于本次金融领域的实例级情感分析任务, 本工作一方面在搜狗细胞词库以及网络信息的爬虫之上, 构建了金融领域词库; 另一方面通过网络爬虫爬取金融领域的相关新闻、评论、百科等书籍文本信息作为预训练语料库, 并对原始 BERT 模型进行了金融领域全词覆盖的预训练处理. 具体的预训练处理过程如下.

(1) 首先利用 MLM 模型随机遮挡金融语料库句子中 20% 的词条, 然后通过模型预测被遮挡的词条是什么. 对随机遮挡的词条采用针对金融领域全词覆盖预训练的方式: ① 80% 的

词向量在输入时被替换为 [MASK], 若词向量所表示的字是金融领域相关名词的一部分, 则同属该词的其他字符也会被相应替换; ② 15% 的词向量被替换为其他词向量, 同理, 若词向量所表示的字是金融领域相关名词的一部分, 则同属该词的其他字符也会被相应替换; ③ 5% 的词向量输入时保持正常. 金融领域全词覆盖 BERT 模型的生成样例如表 1 所示.

表 1　金融领域全词覆盖 BERT 模型的生成样例

Table 1　Examples of BERT-whole word masking model in the financial field

说明	样例
原始文本	小米集团的股价创历史最高, 但软件、硬件商业模式争议从未中断
金融领域分词文本	小米集团 的股价 创 历史 最高, 但 软件、硬件 商业 模式 争议 从未 中断
原始 BERT 模型的生成样例	小 [Mask] 集 [Mask] 的 股 价 创 历 史 最 高, 但 软 件、硬 件 商 [Mask] 模 式 争 议 从 未 中 断
金融领域全词覆盖 BERT 模型的生成样例	[Mask] [Mask] [Mask] [Mask] 的 股 价 创 历 史 最 高, 但 软 件、硬 件 [Mask] [Mask] 模 式 争 议 从 未 中 断

(2) 利用 NSP 模型来预测金融语料库中的下一个句子. 本工作以 50% 的概率将输入的上一段文书句子和下一段文书句子拼接, 另外 50% 的概率是输入上一段文书句子和非下一个随机文书句子的拼接. 该目标任务与上一任务同时进行, 组成多任务预训练.

至此, 实现了针对金融领域全词覆盖的 BERT 预训练. 通过实例可以发现, 该预训练方式可以提取更多的金融领域信息, 更好地解决金融领域评论中语义模糊、关键特征稀疏的问题. 针对情感分析的下游任务, 在输出层加上 Softmax 层对其进行判断. 该模型整体结构如表 1 所示.

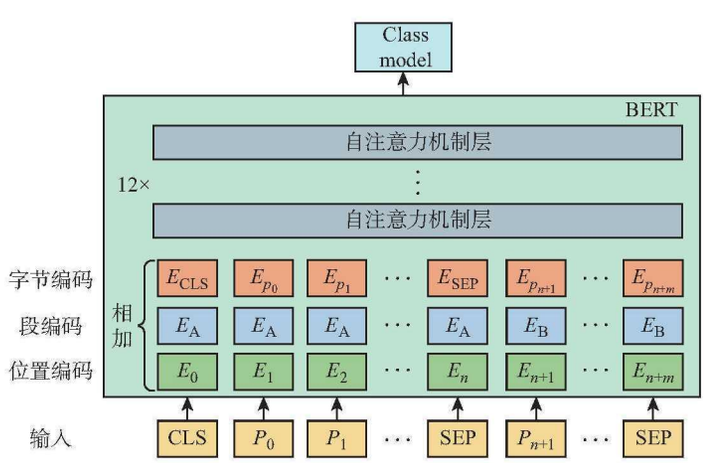

图 1　基于 BERT 的金融领域全词覆盖模型

Fig. 1　Whole word masking based on the BERT model in the financial field

当评论与金融实体输入到基于金融领域全词覆盖的 BERT 预训练模型中时, 模型处理的

计算公式如下:

$$H_{\text{model}} = \text{BERT}(T, E); \tag{1}$$

$$T_{\text{model}} = \text{MLP}(H_{\text{model}}); \tag{2}$$

$$\text{Label} = \text{Softmax}(T_{\text{model}}); \tag{3}$$

式中, MLP 为多层感知器 (multilayer perceptron) 网络, 用于将预训练特征压缩至与类别数同一特征维度, 再通过 Softmax 对特征进行分类.

2.2 基于 BERT 的金融领域特征增强表示模型

为了更好地对金融用户评论中的情感进行提取与分析, 针对金融文本存在的特征提取难、界限相对模糊的问题, 本工作进一步提出了基于 BERT 的金融领域特征增强表示模型.

BERT 预训练模型在训练过程中, 仅仅考虑到了词与词的关系, 通过遮盖部分词或者对下一句话进行预测来学习词向量的特征表示, 并将其压缩至同一特征空间内. 该模型忽略了各个词的词性与命名实体信息, 但显然这些信息是非常重要的. 为此, 本工作将词性特征与命名实体特征进行了精确提取, 与微调 (fine-tune) 后的 BERT 模型, 即添加了 Softmax 层的模型进行融合, 实现更为精确的情感分析. 具体实现步骤如下.

(1) 通过 “结巴分词工具” 对金融领域用户评论进行词性标注处理. 本工作对名词、动词、形容词、副词、介词等 10 类词性进行了标注, 并据此构建词性词汇表. 在训练过程中, 随机初始化和学习词性标注特征 E_{POS}.

(2) 通过 “结巴分词工具” 对金融领域用户评论进行命名实体识别处理. 本工作对人物、时间、地点、金钱、组织这 5 类实体进行了识别. 根据实体类别不同, 进行相应的标注提取, 构建命名实体识别词汇表, 并在训练过程中随机初始化和学习命名实体特征 E_{NER}.

当评论与实体输入到基于 BERT 的金融领域特征增强表示模型时, 得到

$$H_f = \sigma(H_{\text{model}} + M_p E_{\text{POS}} + M_n E_{\text{NER}}), \tag{4}$$

式中: M_p、M_n 为其参数矩阵; σ 为 tanh 激励函数方程.

当基于 BERT 的金融领域特征增强表示模型每层由 Transformer 架构构成时, 该结构不能很好地学习到连续的特征, 即句子中的连续相关性并不能得到充分的学习. 因此, 本工作在模型中加入了双向 LSTM 单元, 以弥补句中顺序特征学习薄弱的问题, 对应的公式为

$$\widehat{H}_t = \text{Bilstm}(H_f). \tag{5}$$

双向 LSTM 网络能够学习词上下文语境长距离的依赖关系, 并根据文本上下文的语境, 输出融合表达的状态向量. 因此, 本工作使用双向 LSTM 隐藏层提取特征增强表示模型输出的词嵌入序列 $H_f = \{h_1, h_2, \cdots, h_n\}$ 中词的特征向量, 学习得到正序词嵌入向量序列 $(h_1, h_2, \cdots, h_n)$ 以及逆序词嵌入向量序列 $(h_1, h_2, \cdots, h_n)$. 以 $\{h_i\}$ 为例, LSTM 隐藏层词嵌入向量计算方法为

$$f_t = \sigma(W_f \cdot [h_{t-1}, q_t] + b_f), \tag{6}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, q_t] + b_i), \tag{7}$$

$$C'_t = \tanh(W_{\mathrm{C}} \cdot [h_{t-1}, q_t] + b_{\mathrm{C}}), \tag{8}$$

$$C_t = f_t * C_{t-1} + i_t * C'_t, \tag{9}$$

$$o_t = \sigma(W_0 \cdot [h_{t-1}, q_t] + b_{\mathrm{o}}), \tag{10}$$

$$h_t = o_t \tanh(C_t), \tag{11}$$

式中: f_t、i_t 和 o_t 分别为遗忘门、输入门和输出门的激活向量, 用来控制 LSTM 单元保存、输入和输出的语义依赖信息; C_t 为细胞 (cell) 状态向量, 代表当前 LSTM 细胞单元存储的信息; C_{t-1} 为上一个 LSTM 细胞状态向量; C'_t 为当前细胞待更新的信息; σ 和 tanh 分别为 sigmoid 和双曲正切函数; h_{t-1} 和 h_t 分别是上一个和本次 LSTM 单元的隐层状态输出. 本工作最终模型的整体结构如图 2 所示.

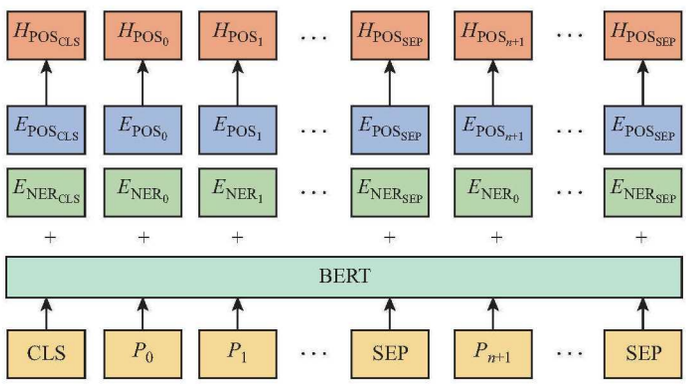

图 2　基于 BERT 的金融领域特征增强表示模型

Fig. 2　Feature enhanced representation based on the BERT model in the financial field

3　实验一

3.1　实验说明

3.1.1　数据集

本实验使用的数据集是 2019 年中国计算机学会 (China Computer Federation, CCF) 举办的 “金融信息负面及主体判定” 比赛中所使用的数据集. 该数据集包含训练集数据量 1 万条, 测试集数据量 1 万条, 主要以金融网络文本 (标题和内容) 和文本中实体列表的形式出现, 具体如表 2 所示.

3.1.2　评估指标

本实验采用负面判定指标 F_1^{s}、主体判定指标 F_1^{e} 和任务整体得分 F_1 共 3 个指标进行评价, 具体公式为

$$P_{\mathrm{s}} = \frac{\mathrm{TP_s}}{\mathrm{TP_s} + \mathrm{FP_s}}, \tag{12}$$

$$R_{\mathrm{s}} = \frac{\mathrm{TP_s}}{\mathrm{TP_s} + \mathrm{FN_s}}, \tag{13}$$

$$F_1^{\mathrm{s}} = \frac{2P_{\mathrm{s}}R_{\mathrm{s}}}{P_{\mathrm{s}} + R_{\mathrm{s}}}, \tag{14}$$

表 2 数据集说明

Table 2 Data set description

字段信息 Field info	类型 Type	描述 Description
id	String	数据 ID Data ID
title	String	文本标题 Text title
text	String	文本内容 Text content
entity	String	给定的实体列表 Given entities
negative	String	文本是否包含负面信息 Does the text contain negative information
key_entity	String	负面主体 Key entity

$$P_{\mathrm{e}}=\frac{\sum_{i=1}^{n}\mathrm{TP}_{ei}}{\sum_{i=1}^{n}\mathrm{TP}_{ei}+\sum_{i=1}^{n}\mathrm{FP}_{ei}},\tag{15}$$

$$R_{\mathrm{e}}=\frac{\sum_{i=1}^{n}\mathrm{TP}_{ei}}{\sum_{i=1}^{n}\mathrm{TP}_{ei}+\sum_{i=1}^{n}\mathrm{FN}_{ei}},\tag{16}$$

$$F_1^{\mathrm{e}}=\frac{2P_{\mathrm{e}}R_{\mathrm{e}}}{P_{\mathrm{e}}+R_{\mathrm{e}}},\tag{17}$$

$$F_1=0.4*F_1^{\mathrm{s}}+0.6*F_1^{\mathrm{e}},\tag{18}$$

式中: 设 n 为金融文本总数; $\mathrm{TP_s}$ 表示负面判定正确的文本数量; $\mathrm{FP_s}$ 表示非负面文本被误判为负面文本的数量; $\mathrm{FN_s}$ 表示负面文本被误判为非负面文本的数量; TP_{ei} 表示第 i 条文本中被正确识别为主体的数量; FN_{ei} 表示第 i 条文本中未被识别出主体的数量; FP_{ei} 表示第 i 条文本中被错误识别为主体的数量.

3.1.3 实验环境与模型参数

本实验的硬件环境配置为 2.20 GHz 的 Intel(R) Xeon(R) 处理器、128 G 内存、32 G 显存的 NVIDIA Tesla V100; 软件环境为 Ubuntu 18.04 操作系统, 开发语言为 Python(v3.7), 采用的深度学习框架为 Pytorch (v1.5).

本实验的参数设置如下: 序列最大长度为 512, 标题加评论最大片段长度为 500, 实体最大长度为 64; Batch 大小为 8, 学习率为 5×10^{-5}, 双向 LSTM 隐藏层大小为 512; 模型的优化算法采用 Adam, 该算法是一种自适应学习率调整的梯度下降算法, 其优点是可以自动调整学习率并且加快收敛速度; Dropout 比率为 0.5, 训练的总轮数为 3 轮.

3.2 对比实验

针对本工作提出的基于 BERT 的金融领域特征增强模型的一种 fine-tune 模式以及特征增强表示, 与 3 种类型的预训练模型进行了对比, 包括原始 BERT, XLNet[15]以及 ERNIE(enhanced representation through knowledge integration).

(1) 原始 BERT. 直接使用未经处理的 BERT 预训练模型, 在模型上加入 MLP 以进行情感识别.

(2) XLNet 是卡内基梅隆大学和谷歌大脑在 2019 年提出的一个新的预训练模型. 执行多任务时的性能超越 BERT. 它是在保留自回归语言模型 (autoregressive language modeling, ALM) 的形式下, 结合了自编码语言模型 (autoencoding language modeling, ALM) 的优势, 提出了排列语言模型 (permutation language modeling, PLM). 同时还基于 Transfomer-XL[16], 有更好的处理长文本的能力.

(3) ERNIE 是百度于 2019 年提出的基于知识增强模型. 通过建模大量数据中的实体概念等先验语义知识, 学习真实世界的语义关系.

3.3　实验结果及分析

3.3.1　对比实验

表 3 为不同模型的对比结果. 可以看到, 本工作提出的最终模型在 F_1^{s}、F_1^{e} 和 F_1 3 个评测指标上分别为 86.38、87.28 与 86.92, 相较于原始 BERT 模型有了大幅的提升, 与另 2 个模型相比也存在很大优势 (在 F_1 指标上比 XLNet 模型高 4.6 个点, 比 ERNIE 模型高 8.8 个点). 这说明本工作提出的基于金融领域全词覆盖与特征增强的情感分析模型对金融评论能够进行精确的分析, 对于情感的抽取更具优势, 在金融情感分析任务上具有更优的效果.

表 3　不同模型之间的对比结果

Table 3　Comparison results between different models　　%

模型名称	F_1^{s}	F_1^{e}	F_1
BERT	78.32	82.42	80.78
XLNet	81.70	82.68	82.29
ERNIE	76.03	79.48	78.10
本工作最终模型	86.38	87.28	86.92

3.3.2　消融实验

本工作主要设计了如下 3 种消融实验方案来证明本工作模型的实验效果: ①金融领域全词覆盖 BERT, 对原始 BERT 预训练模型进行了金融领域全词覆盖的预训练, 并在模型上加入 MLP 以标注答案的位置; ②金融领域全词覆盖 BERT+特征增强表示, 在经过金融领域全词覆盖预训练后的 BERT 模型的基础上, 加入了命名实体特征、词性标注特征, 来增强语言模型的表示与学习能力; ③金融领域全词覆盖 BERT+特征增强表示 + 下游网络, 在基于金融领域全词覆盖与特征增强表示模型的基础上, 加入了双向 LSTM 网络来学习词条之间的序列特征.

本工作在最终模型上进行了消融实验, 用来分析增加的各模块对模型性能表现的影响, 结果如表 4 所示. 可以发现, 模型的每个组件在提高性能方面都起着关键作用.

首先, 采用了原始的 BERT 语言模型, 在模型上加入 MLP 层以在段落中找到答案的位置. 实验结果显示, 其在 F_1^{s}、F_1^{e} 和 F_1 3 个指标上均为最低值.

其次, 对经过金融领域全词覆盖预训练后的 BERT 模型进行了测试. 相对于原始 BERT 语言模型, 在 F_1 结果上有 3.61 个点的提升, 证实了所提出的金融领域全词覆盖预训练的有效性.

接着, 在经过金融领域全词覆盖预训练后的 BERT 模型的基础上, 加入命名实体特征、词性标注特征, 来增强语言模型的表示与学习能力. 与金融领域全词覆盖的 BERT 模型相比, 在

F_1 结果上有 0.90 个点的提升, 证明了特征增强表示对金融领域情感分析任务的有效性.

最后, 加上双向 LSTM 网络来学习词条之间的序列特征, 构成最终模型, 在 F_1 结果上继续提升了 1.63 个点, 证实了本工作提出的基于金融领域全词覆盖与特征增强的情感分析模型的有效性.

表 4 消融实验对比结果

Table 4 Comparison results of ablation experiments %

模型名称	F_1^s	F_1^e	F_1
BERT	78.32	82.42	80.78
金融领域全词覆盖 BERT	83.05	85.28	84.39
金融领域全词覆盖 BERT+ 特征增强表示	84.29	85.96	85.29
本工作最终模型	86.38	87.28	86.92

3.3.3 误差分析

本工作对最终模型进行了误差分析. 结果发现, 当评论长度大于 512 个字符时, 由于 BERT 模型无法一次性提取到评论的全部特征, 因此会因特征缺失而导致模型对情感预测不准确. 为此, 本工作提出的最终模型虽优于其他模型, 但在长文本的情感分析上, 仍需要进一步的改进和优化.

4 实验二

4.1 实验说明

为了进一步证明本工作提出的基于 BERT 的金融文本情感分析模型的有效性, 本工作参考了祝清麟等[17]关于金融领域实体级细粒度情感分析语料库的构建方法, 设计并构建了自己的金融领域细粒度情感分析数据集.

本工作从人民网、新浪财经、东方财富网等各类金融数据网站进行了数据爬取, 通过 Scrapy 框架共爬取了 28 632 条金融新闻与评论. 通过数据清洗与筛选后, 选择了 7 824 条作为本工作的数据集, 并对其进行细粒度的实体、情感标注. 本工作共计标注了如下 3 类情感 (其中实体为加粗字体): ①积极情感 ——**雅迪**在技术方面 20 年如一日的深耕, 以核心技术筑起了坚不可摧的品牌护城河, 奠定了其电动两轮车领域当之无愧的领先者, 实现销量与品牌齐飞; ②中性情感 ——**小米**研发的澎湃 C1 图像信号处理芯片, 可实现更快的自动对焦, 大幅提升在暗光条件下的拍摄效果, 虽然这只是一颗小芯片, 但技术含量很高, 目前国内仅有两家能量产; ③消极情感 —— 短短两年不到的时间, 美国对**华为**进行了四轮制裁, 一轮比一轮狠毒, 把华为消费者业务逼到极端困难, 无法发货.

由于每段新闻或者评论中有可能存在两种以上的金融实体, 因此本工作通过对 7 824 条新闻、评论进行标注后, 共得到 9 251 条数据, 其中积极情感 3 579 条、中性情感 1 758 条、消极情感 3 914 条. 将 9 251 条数据按照情感分类均匀分配, 其中训练集 3 600 条、验证集 2 000 条、测试集 3 651 条.

4.2 对比实验

与实验一相同, 实验二也与原始 BERT、XLNet、ERNIE 模型进行了对比. 与其他情感分析任务相同, 采用了准确率和 Macro-F_1 作为评测标准, 结果如表 5 所示. 可以看出, 本工作提

出的最终模型取得了 77.21% 的准确率和 74.25% 的 Marcro-F_1 值, 均为对比模型中的最优效果, 从而进一步证实了本工作最终模型在金融领域情感分析任务上的优越性.

表 5　对比实验结果

Table 5　Comparison results　　%

模型名称	准确率	Macro-F_1
BERT	75.32	71.52
XLNet	76.95	72.31
ERNIE	73.15	71.36
本工作最终模型	77.21	74.25

4.3　误差分析

为了更好的优化和改进模型, 本工作对实验结果进行了分析, 发现判断错误的情况主要有以下两类.

(1) 数字过多, 导致预训练模型在处理文本时, 文本过长, 关键信息稀疏.

具体示例: 年报显示, 2020 年, 五粮液实现营业收入 573.21 亿元, 同比增长 14.37%, 净利润 199.55 亿元, 同比增长 14.67%; 2021 年一季报显示, 公司一季度实现营业收入 243.25 亿元, 同比增长 20.19%; 净利润 93.24 亿元, 同比增长 21.02%, 业绩整体增速不断提升.

由于以上示例中数字信息过多, 使得具体情感相对稀疏, 从而导致模型将该实体的情感判断为中性, 发生分类错误.

(2) 实体过多, 或实体相近导致模型无法很好地对实体情感进行判断, 将实体情感混淆.

具体示例: 2018 年华夏幸福被爆出资金链危机. 2018 年 9 月和 2019 年 4 月, 中国平安先后斥资 137.7 亿、42.03 亿承接华夏幸福的股权, 助其暂时度过了难关. 同时, 华夏幸福基业控股股份公司进一步增持华夏幸福股权, 并对华夏幸福的未来发展持积极态度.

在以上示例中, “华夏幸福” 的情感应为负面, 但由于文本中实体过多, 且 “华夏幸福”, “华夏幸福基业控股股份公司” 两个实体语义相近, 导致模型对实体情感判断发生错位, 将其情感分为中性, 发生分类错误.

5　结 束 语

本工作针对金融领域文本情感分析任务, 为获取更多的金融文本特征, 提出了基于金融领域的全词覆盖与特征增强的 BERT 预处理模型. 为验证本模型的先进性和有效性, 分别对模型进行了对比实验和消融实验. 结果表明, 本模型在金融文本情感分析任务上取得了更高的分类准确率. 同时也发现, 当评论长度大于 512 个字符时, 由于 BERT 模型无法一次性提取到评论的全部特征, 本模型识别的准确率受限. 因此在以后的研究中, 较长文本的情感识别也是一个重要的研究课题.

参考文献:

[1] Pang B, Lee L. Opinion mining and sentiment analysis [M]. Hanover: Now Publishers Inc, 2008.

[2] Jiao J, Zhou Y. Sentiment polarity analysis based multi-dictionary [J]. Physics Procedia, 2011, 22: 590-596.

[3] Jurek A, Mulvenna M D, Bi Y. Improved lexicon-based sentiment analysis for social media analytics [J]. Security Informatics, 2015, 4(1): 9.

[4] Li F. The information content of forward-looking statements in corporate filings: a naïve Bayesian machine learning approach [J]. Journal of Accounting Research, 2010, 48: 1049-1102.

[5] Hai Z, Cong G, Chang K, et al. Analyzing sentiments in one go: a supervised joint topic modeling approach [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(6): 1172-1185.

[6] Singh J, Singh G, Singh R. Optimization of sentiment analysis using machine learning classifiers [J]. Human-centric Computing and Information Sciences, 2017, 7: 1-32.

[7] Al-Amrani Y, Lazaar M, El-Kadiri K E. Random forest and support vector machine based hybrid approach to sentiment analysis [J]. Procedia Computer Science, 2018, 127: 511-520.

[8] 杨开漠, 吴明芬, 陈涛. 广义文本情感分析综述 [J]. 计算机应用, 2019, 39(S2): 6-14.

[9] Man X, Luo T, Lin J. Financial sentiment analysis (FSA): a survey [C]// 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS). 2019: 617-622.

[10] 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. 2019: 4171-4186.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[12] Matthew E P, Neumann M, Lyyer M, et al. Deep contextualized word representations [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 2227-2237.

[13] Alec R, Karthik N, Tim S, et al. Improving language understanding by generative pre-training [EB/OL]. [2020-12-01]. http://www.nlpir.org/wordpress/wp-content/uploads/2019/06/Improving-language-understanding-by-generative-pre-training.pdf.

[14] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [EB/OL]. [2020-12-01]. https://d4mucfpksywv.cloudfront.net/better-language-odels/language-models.pdf.

[15] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 5753-5763.

[16] Dai, Z H, Yang Z L, Yang Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2978-2988.

[17] 祝清麟, 梁斌, 徐睿峰, 等. 结合金融领域情感词典和注意力机制的细粒度情感分析 [J]. 中文信息学报, 2022, 36(18): 109-117.

(责任编辑: 丁嘉羽)