软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, 2022,33(10):3565–3581 [doi: 10.13328/j.cnki.jos.006630]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 融合知识感知与双重注意力的短文本分类模型*

李博涵[1,2,3], 向宇轩[1], 封 顶[1], 何志超[1,4], 吴佳骏[1], 戴天伦[1], 李 静[1]

[1](南京航空航天大学 计算机科学与技术学院/人工智能学院/软件学院, 江苏 南京 211106)

[2](高安全系统的软件开发与验证技术工业和信息化部重点实验室(南京航空航天大学), 江苏 南京 211106)

[3](软件新技术与产业化协同创新中心, 江苏 南京 211106)

[4](南京市公安局, 江苏 南京 211106)

通信作者: 李博涵, E-mail:bhli@nuaa.edu.cn

**摘 要**: 文本分类任务作为文本挖掘的核心问题, 已成为自然语言处理领域的一个重要课题. 而短文本分类由于稀疏性、实时性和不规范性等特点, 已成为文本分类亟待解决的问题之一. 在某些特定场景, 短文本存在大量隐含语义, 由此给挖掘有限文本内的隐含语义特征等任务带来挑战. 已有的方法对短文本分类主要采用传统机器学习或深度学习算法, 但该类算法的模型构建复杂且工作量大, 效率不高. 此外, 短文本包含有效信息较少且口语化严重, 对模型的特征学习能力要求较高. 针对以上问题, 提出了KAeRCNN模型, 该模型在TextRCNN模型的基础上, 融合了知识感知与双重注意力机制. 知识感知包含了知识图谱实体链接和知识图谱嵌入, 可以引入外部知识以获取语义特征, 同时, 双重注意力机制可以提高模型对短文本中有效信息提取的效率. 实验结果表明, KAeRCNN模型在分类准确度、*F*1值和实际应用效果等方面显著优于传统的机器学习算法. 对算法的性能和适应性进行了验证, 准确率达到95.54%, *F*1值达到0.901, 对比4种传统机器学习算法, 准确率平均提高了约14%, *F*1值提升了约13%. 与TextRCNN相比, KAeRCNN模型在准确性方面提升了约3%. 此外, 与深度学习算法的对比实验结果也说明, 该模型在其他领域的短文本分类中也有较好的表现. 理论和实验结果都证明, 所提出的KAeRCNN模型对短文本分类效果更优.

**关键词**: 短文本分类; 知识图谱; 注意力机制; TextRCNN; 实体消歧

**中图法分类号**: TP18

中文引用格式: 李博涵, 向宇轩, 封顶, 何志超, 吴佳骏, 戴天伦, 李静. 融合知识感知与双重注意力的短文本分类模型. 软件学报, 2022, 33(10): 3565–3581. http://www.jos.org.cn/1000-9825/6630.htm

英文引用格式: Li BH, Xiang YX, Feng D, He ZC, Wu JJ, Dai TL, Li J. Short Text Classification Model Combining Knowledge Aware and Dual Attention. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3565–3581 (in Chinese). http://www.jos.org.cn/1000-9825/6630.htm

## Short Text Classification Model Combining Knowledge Aware and Dual Attention

LI Bo-Han[1,2,3], XIANG Yu-Xuan[1], FENG Ding[1], HE Zhi-Chao[1,4], WU Jia-Jun[1], DAI Tian-Lun[1], LI Jing[1]

[1](College of Computer Science and Technology, College of Artificial Intelligence and College of Software, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

[2](Key Laboratory for Safety-critical Software Development and Verification, Ministry of Industry and Information Technology (Nanjing University of Aeronautics and Astronautics), Nanjing 211016, China)

[3](Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China)

[4](Nanjing Municipal Public Security Bureau, Nanjing 211106, China)

* 基金项目: 国家自然科学基金(62172351, 61728204); 高安全系统的软件开发与验证技术工业和信息化部重点实验室(NJ2018014); 中国学位与研究生教育学会(B-2017Y0904-162); 华为创新 DB IRP (CCF-HUAWEI DBIR2020001A)
本文由“智慧信息系统新技术”专题特约编辑邢春晓研究员、王鑫教授、张勇副研究员、于戈教授推荐.
收稿时间: 2021-07-20; 修改时间: 2021-08-30; 采用时间: 2021-12-24; jos 在线出版时间: 2022-02-22

**Abstract**: As the core problem of text mining, text classification task has become an essential issue in the field of natural language processing. Short text classification is a hot-spot topic, and one of many urgent problems to be solved in text classification due to its sparseness, real-time, and non-standard characteristics. In certain specific scenarios, short texts have many implicit semantics, which brings challenges to tasks such as mining implicit semantic features in limited texts. The existing research methods mainly apply traditional machine learning or deep learning algorithms for short text classification. However, this series of algorithm is complex and requires enormous cost to build an effective model, meanwhile, the algorithms are not efficient. In addition, short text contains less effective information and abundant colloquial language, which requires a stronger feature learning ability of the model. In response to the above problems, the KAeRCNN model is proposed based on the TextRCNN model, which combines knowledge aware and the dual attention mechanism. The knowledge-aware is constructed in two parts, which includes the stage of knowledge graph entity linking and knowledge graph embedding, as external knowledge can be introduced to obtain semantic features. At the same time, the dual attention mechanism can improve the model's efficiency in extracting effective information from short texts. Excessive experimental results show that the KAeRCNN model proposed in this study is significantly better than traditional machine learning algorithms in terms of classification accuracy, the *F*1 score, and practical application effects. The performance and adaptability of the algorithm are further verified with different datasets. The accuracy rate of the proposed approach reaches 95.54%, and the *F*1 score reaches 0.901. Compared with the four traditional machine learning algorithms, the accuracy rate is increased by about 14% on average, and the *F*1 score is increased by about 13%. Compared with TextRCNN, the KAeRCNN model improves accuracy by about 3%. In addition, the experimental results of comparison with deep learning algorithms also show that the proposed model has better performance in classification of short text from other fields. Both theoretical and experimental results indicate that the KAeRCNN model proposed in this study is effective for short text classification.

**Key words**: short text classification; knowledge graph; attention mechanism; TextRCNN; entity disambiguation

文本分类是自然语言处理[1]的一个基本问题, 文本分类应用包罗万象, 如文档组织、新闻过滤、垃圾邮件检测、信息检索、网络舆情发现等, 因此, 文本分类模型具有广泛的应用价值. 特征表示是文本分类中的一个关键问题, 传统的特征表示方法主要采用机器学习, 致力于从文本中选择更具有区分性的特征. 然而, 在对短文本进行分类时, 通常容易受到很多无用词影响, 且由于短文本篇幅短小带来的特征稀疏性, 模型的特征学习能力往往较低, 不满足实际应用场景的需求. 因此, 短文本分类逐渐成为自然语言处理的重要问题.

目前, 短文本分类的内容涉及丰富的话题与信息. 随着微博、微信等成为人们日常生活的主要信息来源, 社交平台中评论与交流的短文本信息呈爆炸式增长. 短文本分类技术被广泛应用于情感分析、垃圾邮件过滤、恶意评论过滤等方面. 相较于长文本, 短文本由于缺乏足够的上下文信息, 且用词不规范、口语化存在大量数据噪声, 导致有限的文本中蕴含了大量的隐含信息; 同时, 在特征稀疏的环境下存在文本语义歧义现象. 如何对短文本进行高精度、高效率的分类, 是目前短文本信息处理中的一个重要问题. 社会各界的各个领域(如公安、新闻等)中有大量亟待解决的短文本分类问题. 因此, 短文本分类问题研究存在广泛的实际应用价值. 各种支持短文本的平台出现, 在给人们带来便利的同时, 也给监管和治理带来了更多挑战, 特别是对公安人员的海量信息处理能力提出了新的要求. 国家多次强调大力发展公安大数据战略. 智慧公安, 数据赋能, 要持续推进数据资源开放共享[2], 把大数据作为推动公安工作创新发展的大引擎、培育战斗力生成新的增长点. 警情文本就是一类典型的稀疏短文本, 其中存在大量口语化、冗余错误、信息缺失等问题. 警情是指社会发生治安、犯罪事件后, 必须由警察出警来维护社会稳定的突发性事件, 或者说是危害公共安全的事件. 各地的"警情"和警察的出警数, 是衡量一个地方治安以及社会稳定的重要因素. 警情文本是指接警员在接线时根据报警人的描述记录下来的文本内容, 一般长度在 5–300 字之间, 其中, 150 字以内的警情达 95%. 过去主要依靠人工分类, 但是, 这种处理方式成本高、效率低, 公安机关需要一个能够快速识别警情类别的方法, 用于提高处理接警信息的效率. 因此, 许多可以自动对警情文本分类分析、快速判断警情类别、对警情进行文本分类的模型被提了出来.

传统的机器学习方式通过词本身的语义获取特征, 实现了文本的自动化分类, 一定程度上提高了文本处理工作效率, 但因此往往也忽略了上下文信息与词相关的外部知识, 所以在处理存在歧义的词语时容易受到误导, 有可能将其理解成截然不同的语义. 然而, 结合上下文的词序或外部知识的引入可以消除这种误导. 同时, 外部知识的引入增强了文本中的特征, 解决了短文本的稀疏性问题, 提高了短文本的分类精度. 融合深度

学习的分类方法, 能够提高文本分类的精度, 但是训练代价也相对提高. 本文在 Yoon Kim[3]和 Yin Wen-Peng[4]工作的基础上, 结合短文本的特点, 基于 TextRCNN 模型[5], 提出了融合知识图谱和双重注意力机制的 KAeRCNN 模型.

KAeRCNN 模型链接外部知识图谱, 通过引入相关实体的知识, 来丰富词的语义信息; 通过双向循环神经网络, 来增加上下文对语义理解的帮助; 通过引入卷积神经网络以降低模型的复杂度, 提高实用性; 利用双重注意力机制, 提高模型学习短文本特征的能力; 通过词过滤算法计算词的贡献度, 来减少文本偏口语化这一特点对模型分类精度的影响. 通过对稀疏短文本的分析, 本文融合注意力机制和知识图谱[6], 在 TextRCNN 的基础上构建了自动的快速判断文本类别的模型, 提高了相关机关对于文本分类的处理效率, 改进和完善了已有的深度学习方法在文本分类方面的性能. 此外, 本文提出的 KAeRCNN 模型在微博短文本情感分析、媒体评论分析等方面具有较好的适用性. 主要贡献如下:

(1) 在对稀疏短文本进行分类时, 将双重注意力机制与 TextRCNN 相结合, 提高了模型特征学习的能力. 将词过滤算法用于短文本分类, 通过计算词贡献度, 过滤掉短文本中出现的高频词和冗余词等, 减少了影响文本分类精确度的干扰项, 降低了某些短文本偏口语化这一特点对模型分类精度的负面影响;

(2) 本文提出的 KAeRCNN 模型通过知识图谱嵌入的方式, 对需要分类的文本中的实体加以增强, 使其具有更多的语义信息, 有效地消除了实体歧义, 并提高了特征表示能力;

(3) 通过理论和实验分析, 将本模型应用于开源新闻数据集以及脱敏后的警情数据集, 在公开短文本与特殊短文本上均提高了分类的准确率和效率. 实验结果表明: 我们的方法在两种数据集上具有较好的表现, 模型具有一定的泛化能力, 大幅降低了人工成本, 具有显著的实际效用.

## 1 相关工作

### 1.1 基于Word2Vec词向量训练

在文本处理过程中, 利用 HanLP 进行文本切词后, 需要转化为词向量后再进行特征提取. 目前, 词向量训练方法使用较多的有 Word2Vec[7]、Glove[8]和 FastText[9]等方法. Word2Vec 是基于分布式词嵌入的方法, 其目的是利用给定的语料库的文档中单词的共现来检测词之间的意义和语义关系, 是一种常见的可以训练词向量的工具. 目前, Word2Vec 在训练词向量时采用如下两种训练模式.

1) Skip Gram: 根据目标单词预测上下文;

2) CBOW (continuous bags of words): 根据上下文预测目标单词, 最后使用模型的部分参数作为词向量.

Skip Gram 在训练数据量较小的情况下运行良好, 甚至可以准确地表示罕见的单词或短语; 而 CBOW 的训练效率优于 Skip Gram, 对频繁出现的单词的预测准确率略高. 针对短文本分类的情况, 本文将 Word2Vec 作为训练词向量的方法, 相较其他方法, Word2Vec 更加快速、有效, 能够满足短文本分类实时性的要求.

### 1.2 基于机器学习的文本分类

传统的机器学习方法处理短文本的过程主要分为 3 个阶段, 分别为文本预处理、文本的特征选择和文本训练.

(1) 文本预处理阶段主要包括文本分词和去停用词. 在处理中文文本时, 分词是首要任务, 一般根据词的个数衡量文本长短, 并且分词的精度对分类效果影响显著. 所以, 一个快速且精准的分词工具非常重要. 目前, 中文分词技术经过多年发展已逐渐成熟, 主流的分词技术主要包括 Jieba、HanLP、Poading Analyzer 等;

(2) 文本特征的选择阶段主要是提取具有代表性和区分性的词语对该文本加以表示, 这一步骤降低了文本特征的空间数量, 提高了文本的分类速度和准确度. 常用的文本特征提取方法有互信息(MI)、

词频-逆向文件频率(TF-IDF)等, 同时会抽取 2-gram 和 3-gram 进行特征统计;

(3) 在文本训练分类阶段, 目前有很多可供选择的传统机器学习方法, 例如常用的逻辑回归(logistic regression, LR)[10]、最近邻算法(*K*-nearest neighbor, KNN)[11]、朴素贝叶斯(naive Bayes, NB)[12]、支持向量机(support vector machines, SVMs)[13]等. 这些算法的特征提取方式不完全相同, 但都依赖情感词典来提取特征. 特征提取对分类的重要性不言而喻, 所以情感词典的构建就至关重要. 不同类型的文本处理需要的情感词典各不相同, 因此, 情感词典不具有普遍适用性. 由于警情文本相对较短、特征不明显、语义稀疏, 无法通过外部语料或附加信息来丰富文本, 故本文不使用传统机器学习方法进行分类.

本文将在第 3.3.4 节将提出的 KAeRCNN 模型从理论和实验两方面, 与 LR、KNN、NB、SVMs 这 4 种传统机器学习算法进行比较.

### 1.3 基于深度学习的文本分类

如今有大量深度学习[14]的方法已经应用到文本分类领域. 2014 年, Kim[3]提出将 TextCNN 模型用于文本分类. 2015 年, Lai 等人[5]在 TextCNN 的基础上将其中的卷积层用 RNN 替换, 提出了 TextRCNN 模型处理文本分类问题. 2018 年, Liu 等人[15]提出了一种 RCNN-HLSTM 的深度分层网络模型进行情感分类. 这些利用深度学习方法来学习文本内容的算法与传统的机器学习方法相比, 可以更好地处理数据稀疏时的特征, 降低人为因素对文本分类精度的影响. 但是, 当将这些方法应用到更加复杂且口语化的警情文本分类场景时, 并不能得到理想的分类效果. 目前有很多优秀的深度神经网络分类算法, 如 DBN[16]、CNN[17]、TextCNN[3]、SDA[18]、RNN[19]、TextRCNN[5]等. 大多数情感分析与文本分类的研究重点为基于上述经典文本表示方法和分类算法进行改进融合, 但融合后分类算法比较复杂, 时间复杂度相对较高, 应用性低. 近年来, 在短文本分类的任务中, 许多方法尝试融入更多复杂的神经网络结构. 2018 年, Zeng 等人[20]用主题记忆网络对主题模型和短文本分类进行联合训练, 通过多任务训练主题模型增强了短文本中的语义信息. 2019 年, Hu 等人[21]提出将异构图神经网络用于短文本分类, 用异构图注意力的方法将异构图嵌入短文本分类中. 2021 年, Zhao 等人[22]使用 GRW 模型和 FastText 模型, 对电信用户投诉短文本进行了分类处理.

CNN 在自然语言处理[23]、计算机视觉、语义分析等领域有着广泛的应用. Zhou 等人[24]对 CNN 在这些领域的研究进展提供了详细介绍. 该研究表明: CNN 在进行局部连接、权值共享和池化操作时, 可以有效地降低计算的复杂度. 在自然语言处理方面, Liu 等人[15]使用区域 CNN 保留句子在评论中的时序关系, 整体提高了其模型效果. 但在偏重口语化的警情文本中, 我们暂未发现应用 CNN 进行该类文本分类的研究.

RNN 在处理序列输入数据上效果突出, 常用于处理文本数据. 该模型考虑了时序因素, 通过逐词分析文本, 并将先前所有文本的语义存储在固定大小的隐藏层中, 能够更好地捕捉上下文信息. 目前, RNN 在用于文本分类任务时, 常常利用其两种改进算法: 长短时记忆(long short time memory, LSTM)[25]和门控循环单元(gate recurrent unit, GRU)[26], 通过引入相应的控制门, 解决了长序列信息经过传统 RNN 处理后保留较多冗余信息以及传统 RNN 训练时出现的梯度爆炸问题. 在训练集充足的情况下, 该模型的表达能力还能够进一步得以提升.

基于传统的机器学习的情感分类算法依赖于建立词典去学习表示特征, 因此, 词典的质量很大程度上影响了文本分类效果. 词典构建会受人为因素影响, 不具有领域通用性, 且人工成本高、效率低, 所以本文使用深度学习方法对短文本进行分类. 基于深度学习的分类方法是目前文本分类研究的热点, 且已取得一定成果, 但是随着文本表示方法和分类算法的改进, 模型的时空复杂度也有一定程度的增加. 虽然方法的改进使得分类准确率有一定提高, 但是模型训练时间会随之增加, 算法复现应用难度较大.

### 1.4 基于预训练模型的文本分类

预训练模型可以通过大型语料库学习通用的语言表征, 避免从零开始训练新模型. 预训练模型降低了训练新模型的成本, 可以更好地理解语句的隐含语义. 预训练模型利用上下文信息, 解决了 Word2Vec 等传统词

向量方法无法解决的一词多义的难题, 也可以用于一般的短文本分类任务. 谷歌公司 2017 年研发的 Transformer 模型[27]通过注意力机制设计了 encoder-decoder 架构, 被广泛应用于自然语言处理领域. 2018 年, Devlin 等人[28]在 Transformer 的基础上提出了 BERT 模型, 该模型由基于 Transformer 机制的 Encoder 和 Decoder 层堆叠而成, 预训练任务分为预测下一句话和随机掩码. 但是由于训练时间长这一局限性, 无法应用于对实时性有要求的文本分类任务. 2020 年, Sun 等人[29]提出 ERNIE 模型, 通过知识增强和递增的多任务学习, 极大地增强了词向量的句法、语法表达能力, 并可同时完成多个任务的输出, 为处理短文本中的隐含语义提供了进一步的帮助.

凭借有效的预训练任务与注意力机制, 预训练模型可以更好地理解语义, 在短文本分类任务上较基于深度学习的方法取得了更高的精度. 然而, 基于预训练模型的文本分类方法往往模型参数巨大、收敛缓慢、训练时间长, 并对硬件的要求较为苛刻, 因此这类方法的使用受到了一定程度的限制, 需要根据实际应用背景来使用. 例如: 在处理训练样本稀少的文本分类任务时, 基于预训练模型的方法效果拔群; 然而在有充足训练样本的场景下, 训练时间长以及硬件要求高使其不适用于许多任务.

### 1.5 知识图谱嵌入

典型的知识图谱由数百万个实体-关系-实体三元组($h$,$r$,$t$)组成, 其中, $h$、$r$ 和 $t$ 分别代表三元组的头部、关系和尾部. 给定知识图谱中的所有三元组, 知识图谱嵌入的目标是学习每个实体和关系的低维表示向量, 保留原始知识图谱的结构信息. 近年来, 基于翻译模型的知识图谱嵌入方法因其简洁的模型和优越的性能而受到广泛关注. 2013 年, 受 Word2Vec 词表示学习的启发, Mikolov 等人[30]提出了一种基于表示学习的方法 TransE. TransE 有效地将语义信息作为学习知识表示的唯一特征, 利用向量空间计算语义关系, 极大地缓解了知识图谱中数据稀疏和传统表示学习方法计算效率低的问题. 2014 年, Wang 等人[31]提出了 TransH 的方法. TransH 通过将实体嵌入而映射到关系超平面中, 并允许实体在涉及不同关系时有不同的表示. 2017 年, Lin 等人[32]提出了 TransR 的方法. TransR 为每个关系引入了一个投影矩阵 $M_r$, 将实体嵌入映射到对应的关系空间. 随后, 其他基于翻译模型的知识图谱嵌入方法相继出现.

### 1.6 注意力机制

注意力(attention)机制最初应用于图像识别领域, 模仿人看图像时, 目光的焦点在不同的物体上移动. 当神经网络对图像或语言进行识别时, 每次集中于部分特征上, 识别更加准确. 在衡量特征的重要性上最直观的方法是计算权重, 因此, Attention 模型的结果在每次识别时, 首先计算每个特征的权重, 然后对特征值加权求和. 权值越大, 该特征对当前识别的贡献就越大. Attention 的主要作用就是让模型从关注全部到关注重点, 将有限资源集中在重点信息上, 从而节约资源, 快速获得最有效的信息. 在自然语言处理领域, Attention 机制被广泛应用: Chai 等人[33]使用 Attention 机制从评论中获取词级别的信息, 用于提升 CNN 对文本中重点信息的关注度; Zhang[34]和 Li 等人[35]通过注意力层确定用户的注意力值, 引入众包和 BERT, 进一步计算在社交网络中的文本的霸凌程度. Attention 机制常常可与 LSTM 模型一起使用, 该组合网络在处理上下文信息时体现出了优异的效果; Xiao 等人[36]提出分类标签同样具有语义这一观点, 使用基于标签语义注意力的方法进行多标签文本分类, 在文档和标签之间共享单词表示, 优于传统的多标签分类模型.

## 2 融合知识感知与双重注意力的短文本分类模型

### 2.1 模型框架

本文主要从降低模型复杂度、减少模型训练时间和降低实现难度的角度, 选择深度学习中的 TextRCNN 算法, 引入外部知识图谱, 并结合注意力机制, 对稀疏短文本展开分类模型的构建、训练和评估的应用研究. 本节将详细介绍本文提出的模型, 受表示学习的启发, 我们提出了由知识图谱实体和上下文嵌入加上词向量嵌入的表示方式, 使文本中的特征词获得更丰富的语义表示. 同时, 我们提出了双重注意力机制, 分别学习文档中词的权重和词上下文时序的权重, 最终通过 TextRCNN 得到当前文档每个标签对应的表示, 模型如图 1

所示. 图 1 将稀疏短文本的分类过程总结为 3 个阶段: 知识感知阶段、双向循环神经网络阶段、池化输出阶段, 双重注意力机制分别添加在知识感知阶段和双向循环神经网络阶段.

如图 1 所示, 左部绿色背景部分是知识感知阶段的框架, 这个阶段主要包括 3 个步骤.

(1) 对需要进行分类的文本作文本切词后, 引入注意力机制和词过滤算法, 去除文本中口语词、停用词的影响, 采用 Word2Vec 词向量训练方法将词转换为词向量, 作为词向量嵌入(word embedding);

(2) 在文本中进行命名实体识别, 随后将识别出的实体利用实体链接技术与知识图谱中存在的实体相关联来消除歧义. 基于这些被识别的实体, 将知识图谱中对应实体映射到对应空间, 并将实体周围 1 跳内的所有实体构建成知识子图, 将子图也映射到空间中, 以上二者统称为知识图谱嵌入(KG embedding);

(3) 将词向量嵌入和知识图谱嵌入进行拼接, 丰富了原本词向量中的语义, 输入下一层的循环神经网络中进行训练.

而在图 1 中间淡黄色背景部分是双向循环神经网络阶段, 用于将当前词结合上下文语境作词义理解, 可以有效降低对短文本分类中的数据稀疏所带来的影响. 首先, 将上个阶段拼接后的向量作为输入, 其中, $c_l$ 为该词的左语境, $c_r$ 为该词的右语境. 通过权重矩阵将隐藏层转换为下一个隐藏层, 并在这里引入另一个注意力机制, 为上下文不同的时序分配不同的权重, 以此来得到更优的特征表示.

池化输出阶段包含图 1 中淡蓝色背景的最大池化层和灰色背景的输出层. 最大池化层将双向循环神经网络层的输出作为输入, 通过池化层可以捕获整个文本中的信息. 与平均池化层不同, 最大池化层用来发现文本中最重要的潜在语义因素. 最后经过 Softmax 函数, 在输出层得到属于每种分类标签的概率.

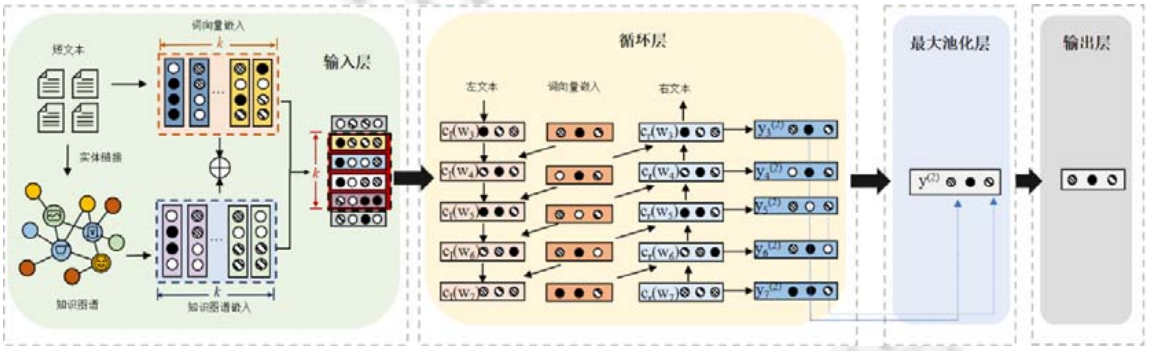


图 1 面向短文本分类的模型框架

### 2.2 引入双重注意力机制

基于神经网络的注意力机制最早被用于视觉图像处理领域. 在文本中抽取词的关键信息, 与人类注意力机制相似. 在阅读短文本时, 通常读者会结合自身的认知, 着重注意某些局部信息, 以快速把握主题. 以警情短文本为例, 近年来, 网络上盛行电信诈骗一类的犯罪, 这类犯罪通常经过缜密的谋划, 案件事实较一般犯罪更为复杂, 而报警人往往因为情绪激动或慌乱紧张, 无法准确描述案件真实情况, 警情文本中常含有大量无意义的语气词以及一些与警情无关的描述. 然而诈骗类案件经常包含‘骗’‘转账’‘金额’‘银行’等特征性词语, 注意力机制能通过分类模型将大部分权重放在这类词上, 以更好地把握主题, 从而避免受到无关因素的干扰. 因此, 本文在模型中引入注意力机制. 例如示例 1 警情.

示例 1: **路 1 号, 理发店外, 报警人和理发店人有纠纷. 今天我陪朋友去理发, 对方拖着我去洗了脸, 我朋友还在店里理发, 事先这个并没有说要收费, 然后现在拦着我不让走, 有纠纷(接警台电话: ****).

此条警情中, 关键词是“纠纷”. 关键词很凸显文本的类别为纠纷求助类, 引入注意力机制可以避免语气词以及与警情无关的冗余描述, 例如示例 1 中的‘我朋友还在店里理发’就是与警情无关的冗余描述, 从而有助于提升文本分类的精度. 本文将注意力机制应用于稀疏短文本分类任务中, 因此, 本文的注意力是指词对该

文本所属类别的注意力. 发现文本中不同注意力, 可以快速捕捉关键信息, 有利于提高文本分类的效率.

注意力模型的本质是一个基于单层神经网络的 Softmax 模型, 它的输入是经过预处理的短文本词向量, 输出是词条对每一类分类标签的贡献度.

设任意输入文本 $W=(w_1,w_2,\ldots,w_n)$, 其中, $n$ 是词的个数, $w_i$ 是输入的词向量. $Y$ 是输出的一维实向量, 表示为$[y^{(1)},y^{(2)},\ldots,y^{(k)}]$, $k$ 为分类类别数, $y^{(k)}$为 $w_i$ 属于 $k$ 类的得分, $Y$ 的计算如公式(1):

$$Y=T\cdot w_i+b \tag{1}$$

其中, $T$ 是权重矩阵, $b$ 是偏置项. 输出 $Y$ 再经 Sigmoid 激活函数及 Softmax 函数, 转换 $w_i$ 属于各类别的概率 $p$. Softmax 的输出如公式(2):

$$p(y=k\mid t_i;T)=\frac{\exp(y^{(r)})}{\sum_{j=1}^{k}\exp(y^{(j)})}(w_i,y_i) \tag{2}$$

在训练阶段, 模型以二元组$(w_i,y_i)$为训练数据, 其中, $y_i$ 是词 $t_i$ 在所处的文本中的类别. 我们用公式(3)作为损失函数来衡量模型预测的性能, 用随机梯度下降法更新参数 $T$:

$$l(T)=\sum_{i=1}^{m}-\log p(y=y_i\mid t_i;T) \tag{3}$$

其中, $m$ 为训练集的词数. 训练结束后, 依次将 $w_i$ 作为输入, 模型输出 $w_i$ 各类别的概率值, 将其作为 $w_i$ 对于类别的注意力向量, 由此能够得到词的注意力矩阵 $M_A$, 见表 1. 其中, 计算每个词对于每一类别的贡献度, 根据词的 $M_A[i]$是 $t_i$ 的注意力向量, 表明 $t_i$ 对各类的置信度.

表 1 词的注意力矩阵 $M_A$

| 注意力向量 | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| $t_1$ | $p(y=1\mid t_1;T)$ | $p(y=2\mid t_1;T)$ | … | $p(y=k\mid t_1;T)$ |
| $t_2$ | $p(y=1\mid t_2;T)$ | $p(y=2\mid t_2;T)$ | … | $p(y=k\mid t_2;T)$ |
| … | … | … | … | … |
| $t_n$ | $p(y=1\mid t_n;T)$ | $p(y=2\mid t_n;T)$ | … | $p(y=k\mid t_n;T)$ |

本文采取双重注意力机制, 分别在输入层和双向循环神经网络层中引入注意力机制. 在分类前, 计算词对各类别的贡献度, 为词过滤做准备; 希望将注意力分配给有实际意义、词性重要的名词或动词, 而相对较少或几乎不分配注意力给介词、语气词、口语词一类的词组, 以此赋予有准确语义的词在文本分类任务中有更高的权重. 在循环神经网络的训练中, 给不同的时序分配不同的权重, 更有效地利用上下文的语义理解当前词的语义; 注意力机制用来捕获短文本中最重要的语义信息, 在双向循环神经网络中等同于捕获最重要的上下文时序.

### 2.3 词过滤算法

词过滤的核心思想是: 在数据进入输入层之前, 根据贡献度大小, 将无用的词过滤掉, 最后将过滤后的数据放入模型中训练. 由于文本多含有一些特定用词, 如警情文本中的“报警”“警察”等, 这些词在对警情分类时作用很小, 并且在预处理阶段的去停用词部分无法去除, 因此, 我们针对该类词语设计了词过滤算法.

注意力向量实际表示文本属于所有分类标签的概率值. 本文将概率作为词对文本的置信度, 概率值越大, 该条文本属于某一标签的置信度越高. 对于概率最大值比较小的词, 其对所有类别的置信度都不高, 这类词对文本分类的贡献度也较弱. 为提高分类效率, 本文在分类前过滤这类词语, 即词过滤. 本文将注意力向量的均方差定义为词的贡献度, 例如词语 $t_i$, 贡献度计算如公式(4):

$$C_{s_i}=\frac{1}{k}\sum_{r=1}^{k}(att_i[r]-a)^2 \tag{4}$$

其中, $att_i[r]$是 $t_i$ 对应类别 $k$ 的注意力; $a$ 是注意力均值, 为 $1/k$. 词过滤算法(算法 1)的 *Delete* 函数表示从文本中删除单词. 该算法需要设置具有启发式特性的超参数 $h$. 在实验中, 我们使用交叉验证来选择 $h$.

**算法 1**. 词过滤算法.

输入: $W=(w_1,w_2,\ldots,w_n)$, $M_A$ \\文本与词的注意力矩阵;
输出: $W'$ \\经过过滤后的文本.
1. $a=1/k$
2. **For each** $w_i$ in $W$
3. $att_i[r]=M_A[i]$
4. $C_{s_i}=\frac{1}{k}\sum_{r=1}^{k}(att_i[r]-a)^2$ \\计算词贡献度
5. **If** $(C_{s_i}<h)$
6. $Delete(W,W_i)$ \\删除置信度低的词
7. **Else**
8. **Return** $C_{s_i}$

### 2.4 知识感知实体增强

知识感知的核心是利用知识图谱发现实体的相关知识, 从而优化实体的特征表示. 为了区分短文本里的知识实体, 我们利用实体链接将短文本中通过命名实体识别发现的实体, 与知识图谱中预定义的实体相关联, 以消除它们的歧义. 例如: 当我们讨论关于“苹果”的信息时, 这里的“苹果”既有可能是一种水果, 也有可能是一种智能手机的品牌. 然而, 单个实体的学习嵌入信息仍然存在局限性, 为了帮助识别实体在知识图谱中的位置并嵌入更多相关信息, 我们将为每个实体嵌入额外的上下文信息. 基于这些被识别出的实体, 我们将从知识图谱中提取该实体的所有关系链接, 将被识别实体 1 跳范围内的其他实体收集起来, 通过实体-关系三元组构建出一个知识子图[37]. 子图中包含对应实体与一跳范围内所有实体, 以及它们之间的关系链接, 构成了实体 $e$ 的上下文信息, 如公式(5):

$$context(e)=\{e_i|(e,r,e_i)\in G \text{ or } (e_i,r,e)\in G\} \tag{5}$$

其中, $r$ 是关系, $G$ 是知识图谱. 由于上下文实体通常在语义和逻辑上与当前实体密切相关, 因此使用上下文可以提供更多的补充信息, 并有助于提高实体的可识别性. 例如: 我们在文本中识别到的实体是“肇事逃逸”, 通过实体链接将其链接到知识图谱中对应的实体. 此时, 我们除了用“肇事逃逸”本身的嵌入来代表实体之外, 还将它的上下文嵌入作为语境, 比如“交通事故”(instance of)、“责任认定书”(相关文件)、“机动车”(车辆类型)、“交警”(所属警类)作为它的标识符. 假定此时我们给定实体 $e$ 的上下文, 上下文嵌入通过公式(6)计算得出:

$$\bar{\boldsymbol{e}}=\frac{1}{|context(e)|}\sum_{e_i\in context(e)}\boldsymbol{e}_i \tag{6}$$

其中, $e_i$ 是通过知识图谱嵌入[38]学习得到的上下文实体嵌入, 此处的上下文嵌入被计算为其上下文实体的平均值. 以上知识图谱[39]相关内容的嵌入经过翻译模型 TransE 得出, 通过分布式表示来描述知识图谱中每个三元组$(h,r,t)$, 将知识图谱中的关系 $r$ 看作实体间的平移向量:

$$l_h+l_r\approx l_t \tag{7}$$

如公式(7), $l_r$ 为关系 $r$ 的向量, $l_h$ 和 $l_t$ 分别为头实体向量和尾实体向量, TransE 模型将 $l_r$ 看作是 $l_h$ 和 $l_t$ 之间的平移, 也可以称为翻译. TransE 定义向量 $l_h+l_r$ 和 $l_t$ 间的距离为 $d$, $S$ 为正确的三元组集合, $S'$为错误的三元组集合, 将损失优化函数$\mathcal{L}$定义为公式(8), 期望正确三元组的距离小, 而错误三元组的距离大:

$$\mathcal{L}=\sum_{(h,l,t)\in S}\sum_{(h',l,t')\in S'}[\gamma+d(\boldsymbol{h}+l,\boldsymbol{t})-d(\boldsymbol{h}'+l,\boldsymbol{t}')]_+ \tag{8}$$

其中, $\gamma$为一个常数, 表示正负样本间的间距; $[x]_+$表示 $\max(0,x)$. 模型训练过程中使用的错误三元组 $S'$由 TransE 将正确三元组 $S$ 中的头实体、关系、尾实体其中之一随机替换为其他实体或关系来生成. 最后, 通过反复训练与调参, 得出一套完整的 TransE 翻译模型, 将知识图谱中的每个实体或关系都能转换为向量表示, 从而实现知识图谱的嵌入.

最终, 知识感知部分的输出由 3 个部分组成: (1) 实体知识图谱实体的嵌入; (2) 该实体 1 跳关系范围内所

有实体嵌入的平均值; (3) 文本中的词经过 Word2Vec 模型训练得到的词向量. 这 3 个部分会被送入下一阶段训练.

### 2.5 TextRCNN分类模型

TextRCNN 分类模型如图 1 右部分所示, 包括 InputLayer(输入层)、BiRNN(双向循环神经网络层)、MaxPool(最大池化层)、Concatenate(拼接层)、Dropout(正则化层)和 Dense(全连接层). TextRCNN 弥补了 TextCNN 在卷积层不容易确定滑动窗口大小的问题, 采用双向循环神经网络[40]取代原来的卷积层. 通过输入层, 获取每个词对应的词向量, 并组合为句子的词向量矩阵, 每个句子的矩阵大小为$(m,k)$, $m$ 为句子的词个数, $k$ 为词向量的维度.

首先, 将词向量矩阵送入双向循环网络层, 得到每个词前向与后向上下文的表示, 分别由公式(9)和公式(10)计算得出:

$$\boldsymbol{c}_l(w_i)=f(W^{(l)}\boldsymbol{c}_l(w_{i-1})+W^{(sl)}\boldsymbol{e}(w_{i-1})) \tag{9}$$

$$\boldsymbol{c}_r(w_i)=f(W^{(r)}\boldsymbol{c}_r(w_{i+1})+W^{(sr)}\boldsymbol{e}(w_{i+1})) \tag{10}$$

其中, $\boldsymbol{c}_l(w_i)$与 $\boldsymbol{c}_r(w_i)$分别为词 $w_i$ 的前向上文表示和后向下文表示, 以公式(11)为例, $w_i$ 表示为输入的第 $i$ 个词; $\boldsymbol{e}(w_{i-1})$是单词 $w_{i-1}$ 的词向量; $\boldsymbol{c}_l(w_{i-1})$表示为当前计算词的上一个词的表示形式; $W^{(l)}$为隐含层的转移矩阵; $W^{(sl)}$是另一个矩阵, 用于将当前词的语义与下一个单词的前向上文表示相结合; $f$ 是一个非线性的激活函数. 由上述两个公式我们可以计算出每个词的前文表示与后文表示. 随后, 通过公式(11)定义出每个词在神经网络中的表示:

$$\boldsymbol{x}_i=[\boldsymbol{c}_l(w_i);\boldsymbol{e}(w_i);\boldsymbol{c}_r(w_i)] \tag{11}$$

$\boldsymbol{x}_i$ 是将词 $w_i$ 的前文表示、词向量、后文表示拼接得到的结果, 再对该结果使用一次 *Sigmoid* 激活函数, 得到的句子表示经过最大池化层后, 输出特征最大的词向量进行 Concatenate 拼接处理, 得到特征向量并迭代送入分类器进行分类. *Binary_crossentropy* 为损失函数, 是二分类的交叉熵, 在处理与本文类似的多标签分类问题时, 其每种分类情况是相互独立、互不干扰的, 因此适于处理某些具有包含关系的标签, 且能同时从属于两种标签. Adam 为算法优化器, 我们充分利用了 Adam 具有自适应学习率的梯度下降算法与动量梯度下降算法的优点, 因此在处理过程中既能适应稀疏梯度, 又能缓解稀疏震荡问题.

### 2.6 基于知识感知的文本分类模型

本文在构建文本分类器时, 在 TextRCNN 模型的基础上, 引入双重注意力机制和知识感知, 一步步改进模型, 最终得到了 KAeRCNN 模型. 在本节, 我们给出 KAeRCNN 模型的进化史, 其模型结构如图 2 所示.

(1) TextRCNN 分类模型. 与第 2.5 节类似, 我们直接利用 TextRCNN 模型对文本进行训练分类, 并将该模型称为 A0_KAeRCNN. 该模型通过卷积学习, 对文本特征进行提取, 进而生成分类模型, 从而更好地学习文本特征, 获得更优的分类效果;

(2) 输入层引入注意力机制. 在模型 A0_KAeRCNN 的基础上, 我们在输入层引入注意力机制, 将改进后的模型称为 A1_KAeRCNN. A1_KAeRCNN 模型计算注意力向量, 并将其作为输入进行双向循环神经网络训练;

(3) 循环神经网络层引入注意力机制. 在 TextRCNN 模型的基础上, 我们在循环神经网络中引入注意力机制, 并将改进后的模型称为 A2_KAeRCNN. 与 A1_KAeRCNN 不同的是, A2_KAeRCNN 模型在循环神经网络中加入注意力机制, 然后将词向量和注意力矩阵作为输入进行双向循环训练;

(4) 输入层和循环神经网络层引入注意力机制. A3_KAeRCNN 模型是在模型 A0_KAeRCNN 的基础上引入双重注意力机制, 即在模型的输入层与循环神经网络层分别引入注意力机制. 该模型在输入层和循环神经网络层上均提高了训练效果;

(5) 融合双重注意力与知识感知. KAeRCNN 模型在 A3_KAeRCNN 模型的基础上, 在输入层引入知识感知, 由实体链接得到文本中实体在知识图谱中对应的实体位置. 当我们定位到知识图谱中的实体

时，如第 2.4 节所述进行知识感知实体增强，将实体嵌入、实体上下文嵌入及词向量嵌入这 3 部分相结合作为输出，然后进行循环训练.

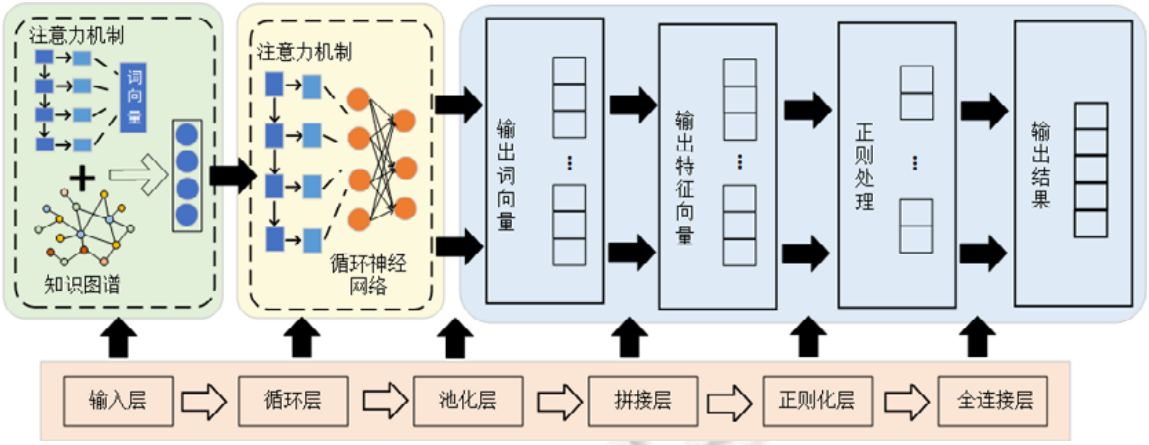


图 2　KAeRCNN 模型结构图

## 3　实验结果及分析

### 3.1　实验数据

本实验分别在公开数据集和私有数据集上进行，在第 3.3 节进行实验，使用精确度和 $F1$ 值作为评估指标验证模型在两类短文本数据上的适用性. 在第 3.3.5 节，选取了部分采用深度学习并具有广泛影响力的文本分类模型作为基线进行对比，以证明本模型较其他方法取得了性能上的提升. 公开数据集是从 THUCNews 中抽取的 20 万条新闻标题，采用的是从新浪新闻 2005–2011 年间的历史数据，文本长度一般在 20–30 之间，共分为 10 个类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐，每个类别约有 2 万条左右的新闻标题. 私有数据集的实验数据来自于某市 2019 年警情数据，提取警情数据中心大数据库中的部分内容，每条数据文本长度约 2–150 字之间，警情类型共分为 9 类：纠纷求助类、盗窃抢夺类、诈骗金融类、人命相关类、黑恶类、赌博类、涉黄涉毒类、自然灾害类. 由于‘报警类型’字段存在大量缺失和错误问题，本文对警情的类型字段进行人工标注. 最后所得警情内容数据集 1 的基本情况见表 2.

表 2　某市 2019 年警情(数据集 1)

| 类别编号 | 类别名称 | >50 字 | >100 字 |
|---|---|---|---|
| 1 | 纠纷求助类 | 1 160 | 44 |
| 2 | 盗窃抢夺类 | 2 115 | 80 |
| 3 | 诈骗金融类 | 9 100 | 1 583 |
| 4 | 人命相关类 | 2 883 | 525 |
| 5 | 黑恶类 | 1 797 | 166 |
| 6 | 赌博类 | 1 142 | 270 |
| 7 | 涉黄涉毒类 | 8 69 | 79 |
| 8 | 自然灾害类 | 6 760 | 3 897 |
| 9 | 其他 | 5 214 | 2 991 |
| 总计\占比 | – | 31040\34.5% | 9635\10.7% |

最后所得警情内容数据集 1 见表 2. 从表 2 中可以看出：90 000 条数据中，有 34.5%的警情大于 50 字，其中涉黄涉毒类大于 50 字警情最少，诈骗金融类最多；有 10.7%的警情大于 100 字，纠纷求助类、盗窃抢夺类、涉黄涉毒类警情分别只有 44 条、80 条、79 条，黑恶类与赌博类警情低于 300 条.

为了对比分类器学习效果，本研究提取纠纷求助类、盗窃抢夺类、涉黄涉毒类、黑恶类与赌博类中报警内容大于 50 字的数据各 1 万条，对诈骗金融类、人命相关类、自然灾害类、其他类中报警内容大于 100 字的

数据各 1 万条, 得到较长文本的数据集 2 的基本情况见表 3.

表 3 警情文本长度大于 50 字(数据集 2)

| 类别编号 | 类别名称 | 较长文本占比(%) | 文本长度(字) |
|---|---|---|---|
| 1 | 纠纷求助类 | 3.7 | >50 |
| 2 | 盗窃抢夺类 | 3.7 | >50 |
| 3 | 诈骗金融类 | 17.3 | >100 |
| 4 | 人命相关类 | 18.2 | >100 |
| 5 | 黑恶类 | 9.2 | >50 |
| 6 | 赌博类 | 23.6 | >50 |
| 7 | 涉黄涉毒类 | 9.0 | >50 |
| 8 | 自然灾害类 | 57.6 | >100 |
| 9 | 其他 | 57.4 | >100 |

### 3.2 实验设置与参数选择

我们使用了第 2.6 节中提出的 5 个模型进行比较, 对于数据集 1、数据集 2, 各分为 9 类, 每类 1 万条, 数据集拆分为训练集和测试集, 比例为 4:1. 使用 5 个模型对数据集 1、数据集 2 训练, 从横向和纵向分别对比相同长度文本不同模型训练效果、不同长度文本相同模型训练效果. 通过 5 个模型分类正确率的情况, 结合 *F*1 得分, 判断模型训练精确度. 模型训练精确度应与分类正确率成正比, 因此在以下实验中, 通过分类正确率的增长来反映模型训练精确度的提升.

本文在输入层引入词过滤算法, 具体词过滤算法已在第 2.3 节中说明, 计算进入输入层前的文本中各个词的贡献度. *h* 值在这里定义为词过滤的阈值, 当词的贡献度小于 *h* 时舍弃该词, 大于 *h* 值时保留该词. 词过滤算法用于去除文本中贡献度较小的词语, 或者在较多的文本中都出现的词语.

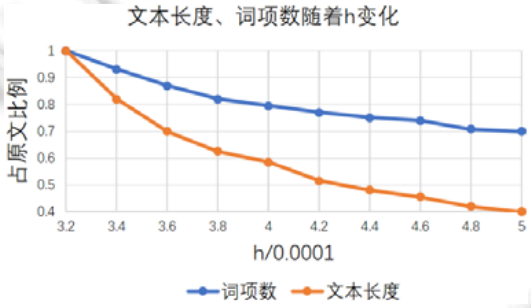


图 3 文本长度随 *h* 的变化情况

图 3 是用第 2.3 节的词过滤算法过滤后的文本长度、词项数占过滤前的比例随 *h* 的变化情况, 其中, 长度是文本的字符数, 词项数是文本中的不重复词数. 本文使用五折交叉验证法, 记录 *h* 过滤的效果关系. 结果表明: 当 *h* 大于 5.0×0.0001 时, 分类正确率低于 70%; 当 *h* 小于 3.1×0.0001 时, 过滤句子数趋于 0. 因此, 本文记录 *h* 的测试区间为[3.2×0.0001,5.2×0.0001]. 从图 3 可知: 当 *h* 为 3.2×0.0001 时, 未过滤任何一个词; 当 *h* 为 5.0×0.0001 时, 词数保留量为 70.41%. 这表明, 本方法过滤了文本中较多高频词, 当文本长度减少时, 仍然能够很好地保留文本词项. 这与一般文本过滤方法结果一致, 未出现短句与长句被集中过滤的情况.

### 3.3 实验结果与分析

#### 3.3.1 *h* 取不同值时对结果的影响

本节比较模型 A0_KAeRCNN 与 KAeRCNN 模型在使用词过滤算法后得到的精确度变化情况. 实验使用的数据集为数据集 2, 得到的正确率变化趋势线如图 4 所示. 从图 4 可以看出: 当 *h*=3.4×0.0001 时, KAeRCNN 模型得到的准确率最高; 当 *h* 值逐渐大于 3.4×0.0001 时, 分类的精确度逐渐降低. A0_KAeRCNN 结果与 KAeRCNN 结果类似. 为了得到更好的训练效果, 应当选取使分类正确率最高的 *h* 值. 我们选取 *h*=3.4×0.0001

进行后续实验.

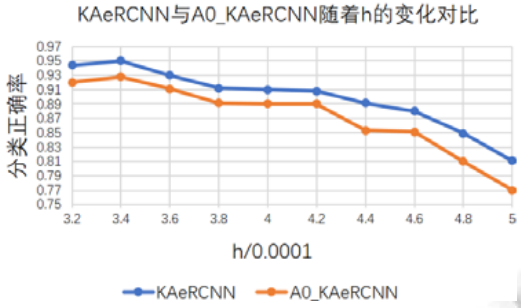


图 4 $h$ 值对 KAeRCNN 与 A0_KAeRCNN 的正确率的影响

3.3.2 模型训练精确度的对比与分析

本实验为成长型实验，使用的数据集为数据集 1 和数据集 2. 我们比较了第 2.6 节中提到的 5 个成长型模型的精确度随着迭代次数的增加产生的变化，最终得到的变化情况如图 5 和图 6 所示.

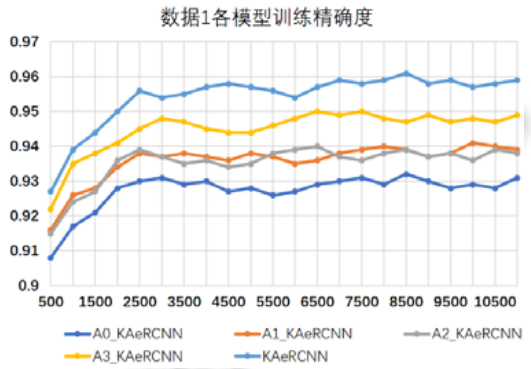


图 5 模型精度随迭代次数变化示意图(数据集 1)

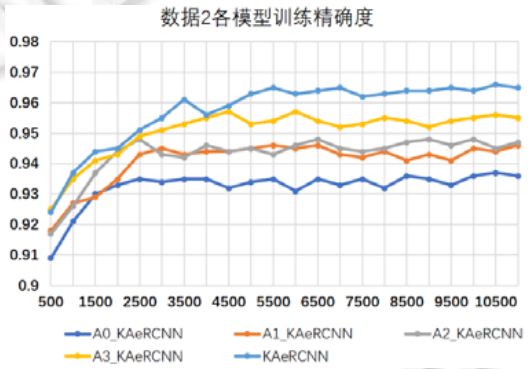


图 6 模型精度随迭代次数变化示意图(数据集 2)

使用数据集 1 进行实验的准确度如图 5 所示, 5 个模型训练效果不一，其中, A0_KAeRCNN 的精确度相对较低，最终稳定在 0.929 左右；A1_KAeRCNN 与 A2_KAeRCNN 的精确度较为接近，最终稳定在 0.937 左右；A3_KAeRCNN 最终的精确度达到 0.949；而 KAeRCNN 最终的精度能达到 0.959. 实验结果表明：在加入双重注意力机制和词过滤后，模型精度从最初的 0.929 提升至 0.959，提升约 3%.

使用数据集 2 进行训练时，各个模型的精度较之前均有提升. 如图 6 所示：A0_KAeRCNN 模型的精确度提升至 0.935 左右；而分类效果最好的 KAeRCNN 模型，其精确度达到 0.964，模型最终效果提升了 3.9%. 由此我们得到以下结论：(1) 使用知识图谱嵌入消除了实体歧义的问题，通过文本外的知识增强了对实体的理解，因此知识图谱嵌入提高了模型的训练效果；(2) 数据集 2 的最终训练效果优于数据集 1 的训练效果，因为较长的文本中实体更多，能够引入更多实体的嵌入，而且更多的时序有利于更多的上下文，因此在短文本分类时，文本长度与精度紧密相关.

3.3.3 $F1$ 值与用时对比

在分类器模型的评价中，常采用 $F1$ 值. $F1$ 值是 precision 和 recall 的调和均值，能够较好地反映神经网络在训练过程中的表现，它的计算方法如公式(12):

$$F1=\frac{2\times P\times R}{P+R} \tag{12}$$

其中, $P$ 为模型的准确度(precision), $R$ 代表模型的召回率(recall). 本文通过计算 $F1$ 值来测试模型在警情文本分类上的表现.

为了验证模型的分类性能, 本实验另随机抽取 4 000 条警情数据作为测试集, 并使用经过数据集 1 训练好的 5 个模型, 分别计算每个模型的 *F*1 值, 最终实验结果如图 7 所示.

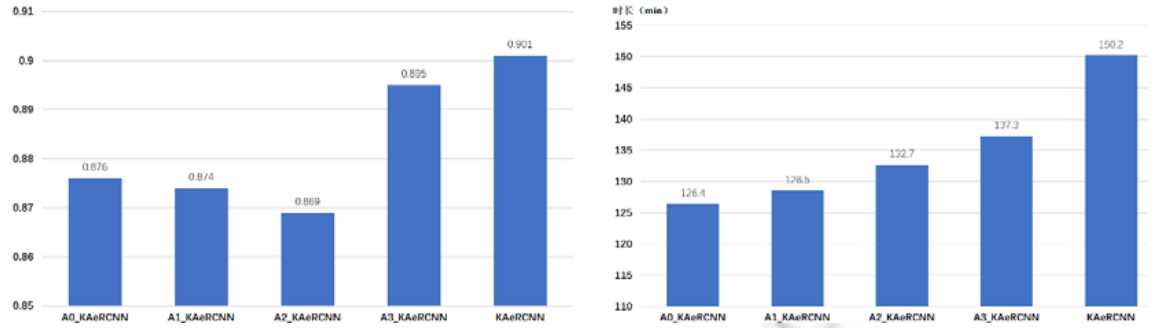


图 7 5 个模型在 *F*1 值上的表现(左)与用时比较(右)

如图 7 所示, 5 个模型的 *F*1 值比较接近, 说明模型相对稳定. 模型 A2_KAeRCNN 的 *F*1 值最低, 为 0.869 (该模型召回率较低, 导致 *F*1 值最低). KAeRCNN 模型 *F*1 值达到 0.901. 综合来看, KAeRCNN 模型表现最佳.

实验用时是考量模型复杂程度的另一个维度, 短文本分类在某些应用场景下往往数据量庞大, 需要在合理而有效的时间内对短文本进行分类. 本实验使用的 5 个模型用时如图 7 所示: 随着模型逐渐复杂, 模型的用时也逐渐增加. 经对比发现, 知识图谱实体链接的用时较多, 并在进行知识图谱嵌入时带来了大量的语义信息. 因此, KAeRCNN 的用时最多. 对比模型 A0_KAeRCNN 与模型 KAeRCNN 可见, 添加知识感知与双重注意力机制在原本模型的基础上增加了约 20%的用时. 对比测试集规模与所用时长, 我们认为, KAeRCNN 模型在短文本分类任务上的效果仍然能够满足实际应用需求.

3.3.4 与传统短文本多分类模型对比

由于警情短文本多分类的文章较少, 我们选取几个传统的机器学习模型进行比较.

通过对各个模型的 Precision、Recall、*F*1 和 Accuracy 进行对比, SVM、KNN、LR、NB 的训练结果可见表 4. 由表 4 可知, NB 模型表现最优, 但其准确度只有 84.07%, 而 *F*1 值为 80.93%, 均低于 KAeRCNN 模型.

表 4 机器学习文本分类方法对比

| | SVM | KNN | LR | NB |
|---|---|---|---|---|
| *F*1 | 0.778 6 | 0.672 7 | 0.798 | 0.809 3 |
| Recall | 0.781 2 | 0.696 2 | 0.795 | 0.805 6 |
| Precision | 0.787 6 | 0.704 3 | 0.803 4 | 0.818 |
| Accuracy | 0.816 9 | 0.751 5 | 0.83 | 0.840 7 |

3.3.5 与深度学习文本分类模型对比

由于警情短文本的特殊性, 与之相关的实验不仅少而且采用的数据集也存在差异, 不具有对比参考的价值. 因此, 本实验采用 KAeRCNN 模型, 在第 3.1 节提及的公开数据集上进行训练. 公开数据集是从 THUCNews 中抽取的 20 万条新闻标题, 采用的是从新浪新闻 2005–2011 年的历史数据, 文本长度普遍在 20–30 之间, 共分为 10 个类别: 财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐, 每个类别约有 2 万条的新闻标题. 数据集以 4:1 的比例拆分成训练集和测试集, 与 TextRNN[41]、TextCNN[3]、TextRNN+Attention[42]、DPRCNN[43]、TextRCNN[5]、FastText[9]、BERT[28]进行对比实验, 结果见表 5.

由表 5 可知: 在基于神经网络的文本分类模型中, FastText 模型的性能最优. 本文提出的基于知识感知与双重注意力的短文本分类模型 KAeRCNN, 由于其在神经网络的基础上, 通过知识图谱引入外部知识帮助训练, 因此在分类准确率上优于对比的深度学习文本分类模型, 主要提升在于对文本中关键词的特征理解能力.

表 5 深度学习文本分类模型对比

| 模型 | 准确率(%) |
|---|---|
| TextRNN | 91.32 |
| TextCNN | 91.72 |
| TextRNN+Attention | 92.10 |
| DPCNN | 92.35 |
| TextRCNN | 92.91 |
| FastText | 93.83 |
| BERT | 94.62 |
| **KAeRCNN** | **95.54** |

3.3.6 与短文本分类模型功能对比

近年来, 针对短文本分类特征稀疏、隐含语义丰富等特点, 较为流行的方法主要侧重于主题模型、图神经网络、多模型融合等改进方面. 本节从使用的方法及其优势与不足等方面, 将 KAeRCNN 模型与近几年短文本分类模型进行比较. 选取对比的模型分别为: 2018 年, Zeng 等人提出的主题记忆网络方法; 2019 年, Hu 等人提出的异构图神经网络方法; 以及 2021 年, Zhao 等人提出的 GRW+FastText 模型. 这 3 种模型分别为主题模型、图神经网络、多模型融合方面的代表, 我们从功能上对近年来出现的几种短文本分类模型进行了对比, 由此也帮助我们在未来能够更好地完善我们的方法. 对比结果见表 6.

表 6 短文本分类模型功能对比

| | 增强语义 | 特征提取能力 | 泛化性 | 分类用时 |
|---|---|---|---|---|
| 主题记忆网络[20] | 无 | 强 | 强 | 长 |
| 异构图神经网络[21] | 较强 | 较强 | 较强 | 长 |
| GRW+FastText[22] | 无 | 强 | 弱 | 较短 |
| KAeRCNN | 强 | 较强 | 较强 | 短 |

由表 6 可知: 异构图神经网络与 KAeRCNN 均通过知识图谱引入了外部知识, KAeRCNN 较异构图神经网络引入实体 1 跳范围内上下文, 因此在增强语义能力方面, KAeRCNN 强于异构图神经网络. 在特征提取方面, 异构图神经网络与 KAeRCNN 依靠注意力机制来增强特征提取能力, 而主题记忆网络和 GRW+FastText 方法分别通过训练主题模型与双重模型叠加的方式增强特征提取能力. 主题记忆网络凭借记忆网络主题推理能力, 能够编码潜在的主题表示, 因此具有强大的泛化能力. 本文提出的 KAeRCNN 通过知识图谱增加了泛化性, 但是, 如果某特定领域不存在相关知识图谱, 则受到限制. 最后, 基于每个模型的复杂程度, 对短文本的分类用时有所不同, KAeRCNN 更多考虑到短文本的实时性, 在处理短文本时相比其他模型更高效.

总结以上对比, 本文提出的 KAeRCNN 模型在语义增强和分类用时方面领先于近几年的短文本分类模型, 在特征提取能力和泛化性方面也有较一般模型更优的能力. 结合以上分析, 我们提出的 KAeRCNN 在模型的特征提取能力与泛化性上还存在改进空间, 未来可进一步加以完善.

## 4 总 结

本文针对短文本中数据稀疏的特征提出了 KAeRCNN 模型. 该模型使用 Word2Vec 训练词向量, 通过知识图谱嵌入引入外部知识以辅助词向量生成, 分别在 TextRCNN 的输入层和双向循环神经网络层上引入注意力机制. 针对短文本偏口语化的特点, 本研究引入词过滤算法, 通过计算词的贡献度, 以对短文本进行词过滤. KAeRCNN 的主要优点表现在以下几个方面: (1) 该模型能够更好地学习短文本特征, 克服传统学习算法无法解决的特征稀疏问题; (2) 利用知识图谱消除词的歧义, 同时增加外部知识来增强对词的理解, 进一步规避了错误并提高了分类精度; (3) 引入双重注意力机制, 用以提高模型的可解释性, 提升了模型特征的捕获能力; (4) 使用词的贡献度过滤算法, 对文本作进一步处理, 提升了文本精度. 对于实验中采用的警情短文本, 由于警情文本长度短, 分词精度相对正规文本较低, 且警情文本涉及的信息种类较多, 未来我们打算对警情文本分词进行优化, 以提高关键词的分词准确率, 进一步增加警情类别, 提升模型质量. 本文使用公开的数据集来测试模型的适用性, 模型既能在特殊的警情文本中取得较好的结果, 也能广泛适用于类似新闻短文本的分类.

**References**:


[1] Wang NY, Ye YX, Liu L, Feng LZ, Bao T, Peng T. Language models based on deep learning: A review. Ruan Jian Xue Bao/ Journal of Software, 2021, 32(4): 1082−1115 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[2] Wang N. Analysis of public security education in the era of big data—Comment on “Smart public security—Policing mode in the era of big data”. Journal of the Chinese Society of Education, 2020, 12: 110−111 (in Chinese with English abstract).

[3] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014. 1746−1751.

[4] Yin WP, Schütze H, Xiang B, Zhou BW. ABCNN: Attention-based convolution neural network for modeling sentence pairs. Trans. of the Association for Computational Linguistics, 2016, 4: 259−272.

[5] Lai SW, Xu LH, Liu J. Recurrent convolutional neural networks for text classification. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. Austin: AAAI, 2015. 2267−2273.

[6] Wang X, Zou L, Wang CK, Peng Y, Feng ZY. Research on knowledge graph data management: A survey. Ruan Jian Xue Bao/ Journal of Software, 2019, 30(7): 2139−2174 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[7] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv Preprint arXiv: 1301. 3781, 2013.

[8] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532−1543.

[9] Sahin A, Hurtado Grooscors H, Góngora-Cortés J. Review of FastTest: A platform for adaptive testing. Measurement: Interdisciplinary Research and Perspectives, 2018, 16(4): 256−263.

[10] Pranckevicius T, Marcinkevicius V. Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In: Proc. of the Advances in Information, Electronic & Electrical Engineering. Vilnius: IEEE, 2017. 1−5.

[11] Abeywickrama T, Cheema M, Taniar D. *k*-nearest neighbors on road networks: A journey in experimentation and in-memory implementation. Proc. of the VLDB Endowment, 2016, 9(6): 492−503.

[12] Mccallum A, Nigam K. A comparison of event models for Naive Bayes text classification. In: Proc. of the AAAI’98 Workshop on Learning for Text Categorization. 1998. 41−48.

[13] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proc. of the Conf. on Machine Learning. Berlin: Springer, 1998. 137−142.

[14] Cheng KY, Wang N, Shi WX, Zhan YZ. Research advances in the interpretability of deep learning. Journal of Computer Research and Development, 2020, 57(6): 1208−1217 (in Chinese with English abstract).

[15] Liu Q, Liang B, Xu J, Zhou Q. A deep hierarchical network model based on sentiment analysis. Chinese Journal of Computers, 2018, 41(12): 2637−2652 (in Chinese with English abstract).

[16] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504−507.

[17] Le-Cun Y, Bottou L, Bengio L, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278−2324.

[18] Vincent P, Larochelle H, Bengio Y, Manzagol P. Extracting and composing robust features with denoising autoencoders. In: Proc. of the 25th Int’l Conf. on Machine Learning. Helsinki: ACM, 2008. 1096−1103.

[19] Socher R, Huval B, Manning C, Ng AY. Semantic compositionality through recursive matrix-vector spaces. In: Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2012. 1201−1211.

[20] Zeng J, Jing L, Yan S, Gao C, Lyu M, King I. Topic memory networks for short text classification. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2018). 2018. 3120−3131.

[21] Hu LM, Yang TC, Shi C, Ji HY, Li XL. Heterogeneous graph attention networks for semi-supervised short text classification. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int’l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hongkong: Association for Computational Linguistics, 2019. 4821−4830.

[22] Zhao J, Yang XJ. Application on text classification of telecom user complaints based on GRW and FastText model. Telecommunications Science, 2021, 37(6): 125−131 (in Chinese with English abstract).

[23] Wang JX, Wang ZY, Tian X. Review of natural scene text detection and recognition based on deep learning. Ruan Jian Xue Bao/ Journal of Software, 2020, 31(5): 1465−1496 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5988.htm [doi: 10.13328/j.cnki.jos.005988]

[24] Zhou FY, Jin LP, Dong J. Review of convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1229−1251 (in Chinese with English abstract).

[25] Zhang SY, Yang Y, Xiao J, Liu XM, Yang Y, Xie D, Zhuang YT. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Trans. on Multimedia, 2018, 20(9): 2330−2343.

[26] Cho K, Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1724−1734.

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000−6010.

[28] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, ed. Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171−4186.

[29] Sun Y, Wang SH, Li YK, Feng SK, Tian H, Wu H, Wang HF. Ernie 2.0: A continual pre-training framework for language understanding. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 8968−8975.

[30] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed rep-resentations of words and phrases and their compositionality. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111−3119.

[31] Wang Z, Zhang JW, Feng JL, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Québec City: AAAI, 2014. 1112−1119.

[32] Lin Y, Liu ZY, Sun MS, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. Austin: AAAI, 2015. 2181−2187.

[33] Chai YM, Yun WL, Wang LM, Liu Z. A cross-domain recommendation model base on dual attention mechanism and transfer learning. Chinese Journal of Computers, 2020, 43(10): 1924−1942 (in Chinese with English abstract).

[34] Zhang AM, Li BH, Wang WH, Wan S, Chen WT. MII: A novel text classification model combining deep active learning with BERT. Computers Materials and Continua, 2020, 63(3): 1499−1514.

[35] Li BH, Zhang AM, Chen WT, Yin HL, Cai K. Active cross-query learning: A reliable labeling mechanism via crowdsourcing for smart surveillance. Computer Communications, 2020, 152: 149−154.

[36] Xiao L, Chen BL, Huang X, Liu HF, Jing LP, Yu J. Multi-label text classification method based on label semantic information. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1079−1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

[37] Wang H, Zhang F, Xie X, *et al*. DKN: Deep knowledge-aware network for news recommendation. In: Proc. of the 2018 World Wide Web Conf. Lyon: Int'l World Wide Web Conf. on Steering Committee, 2018. 1835−1844.

[38] Wang XZ, Gao TY, Zhu ZC, Zhang ZY, Liu XY, Li JZ, Tang J. KEPLER: A unified model for knowledge embedding and pre-trained language representation. Trans. of the Association for Computational Linguistics, 2021, 9: 176−194.

[39] Liu ZY, Sun MS, Lin YK, Xie RB. Knowledge representation learning: A review. Journal of Computer Research and Development, 2016, 53(2): 247−261 (in Chinese with English abstract).

[40] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT, 2014. 2204−2212.

[41] Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: AAAI, 2016. 2873−2879.

[42] Zhou P, Shi W, Tian J, Qi ZY, Li BC, Hao HW, Xu B. Attention-based bidirectional long short-term memory networks for relation classification. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 207–212.

[43] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 562–570.

## 附中文参考文献:

[1] 王乃钰, 叶育鑫, 刘露, 凤丽洲, 包铁, 彭涛. 基于深度学习的语言模型研究进展. 软件学报, 2021, 32(4): 1082–1115. http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[2] 王楠. 浅析大数据时代下的公安教育——评《智慧公安——大数据时代的警务模式》. 中国教育学刊, 2020, 12: 110–111.

[6] 王鑫, 邹磊, 王朝坤, 彭鹏, 冯志勇. 知识图谱数据管理研究综述. 软件学报, 2019, 30(7): 2139–2174. http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[14] 成科扬, 王宁, 师文喜, 詹永照. 深度学习可解释性研究进展. 计算机研究与发展, 2020, 57(6): 1208–1217.

[15] 刘全, 梁斌, 徐进, 周倩. 一种用于基于方面情感分析的深度分层网络模型. 计算机学报, 2018, 41(12): 2637–2652.

[22] 赵进, 杨小军. 基于 GRW 和 FastText 模型的电信用户投诉文本分类应用. 电信科学, 2021, 37(6): 125–131.

[23] 王建新, 王子亚, 田萱. 基于深度学习的自然场景文本检测与识别综述. 软件学报, 2020, 31(5): 1465–1496. http://www.jos.org.cn/1000-9825/5988.htm [doi: 10.13328/j.cnki.jos.005988]

[24] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229–1251.

[33] 柴玉梅, 员武莲, 王黎明, 刘箴. 基于双注意力机制和迁移学习的跨领域推荐模型. 计算机学报, 2020, 43(10): 1924–1942.

[36] 肖琳, 陈博理, 黄鑫, 刘华锋, 景丽萍, 于剑. 基于标签语义注意力的多标签文本分类. 软件学报, 2020, 31(4): 1079–1089. http://www.jos.org.cn/1000-9825/5923.htm [doi: 10.13328/j.cnki.jos.005923]

[39] 刘知远, 孙茂松, 林衍凯, 谢若冰. 知识表示学习研究进展. 计算机研究与发展, 2016, 53(2): 247–261.

**李博涵**(1979－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为时空数据库, 知识图谱, 自然语言处理, 推荐系统.

**吴佳骏**(1998－), 男, 硕士生, CCF 学生会员, 主要研究领域为知识图谱, 表示学习.

**向宇轩**(1998－), 男, 硕士生, CCF 学生会员, 主要研究领域为自然语言处理, 知识图谱.

**戴天伦**(1998－), 男, 硕士生, CCF 学生会员, 主要研究领域为时空数据, 路径规划.

**封顶**(1994－), 男, 硕士, 主要研究领域为自然语言处理.

**李静**(1976－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为数据挖掘, 图像处理.

**何志超**(1990－), 男, 硕士, 主要研究领域为数据挖掘.