

基于注意力机制和 BiGRU 融合的文本情感分析

杨青 张亚文 朱丽 吴涛

人工智能与智慧学习湖北省重点实验室 武汉 430079

华中师范大学计算机学院 武汉 430079

国家语言资源监测与研究网络媒体中心 武汉 430079

(yangqing@mail.ccnu.edu.cn)

摘要 针对简单的神经网络缺乏捕获文本上下文语义和提取文本内重要信息的能力,设计了一种注意力机制和门控单元(GRU)融合的情感分析模型 FFA-BiAGRU。首先,对文本进行预处理,通过 GloVe 进行词向量化,降低向量空间维度;然后,将注意力机制与门控单元的更新门融合以形成混合模型,使其能提取文本特征中的重要信息;最后,通过强制向前注意力机制进一步提取文本特征,再由 softmax 分类器进行分类。在公开数据集上进行实验,结果证明该算法能有效提高情感分析的性能。

关键词: 注意力机制;门控单元;GloVe 词向量;情感分析

中图法分类号 TP399

Text Sentiment Analysis Based on Fusion of Attention Mechanism and BiGRU

YANG Qing, ZHANG Ya-wen, ZHU Li and WU Tao

Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Wuhan 430079, China

School of Computer, Central China Normal University, Wuhan 430079, China

National Language Resources Monitor & Research Center for Network Media, Wuhan 430079, China

Abstract Aiming at the lack of the ability of simple neural networks to capture the contextual semantics of texts and extract important information in texts, a sentiment analysis model FFA-BiAGRU is proposed, which integrates attention mechanism and GRU. First, we pre-process the text and vectorize the words through GloVe to reduce the vector space dimension. Then, through a hybrid model that fuses the attention mechanism with the update gate of the gating unit, it can extract important information in the text features. Finally, the text features are further extracted through the forced forward attention mechanism, and then classified by the softmax classifier. Experiments on public data sets show that the algorithm can effectively improve the sentiment analysis performance.

Keywords Attention mechanism, GRU, GloVe word vector, Emotion analysis

1 引言

情感分析(sentiment analysis)主要是对文本进行分析、处理,抽取出文本中的情感色彩的过程,其实质为一种文本分类。一般的情感分析方法分为基于情感词典的情感分析方法和基于机器学习的情感分析方法。Jha 等^[1]通过构建全自动印度语多领域情感词典来解决目标领域中未标记评论的情感分类问题,以减少人工标记。Xu 等^[2]使用 SVM 模型对航天事件相关内容进行分析,从而掌握航天事件对舆论的影响。但是,这些方法都需要人工标注,在如今大量数据需要处理的情况下,存在一定的弊端。随着深度学习走进人们的视野,越来越多的研究者开始将其应用到不同的领域,目前深度学习

在机器视觉、语音识别、自然语言处理等领域都取得了优秀的成果。卷积神经网络^[3](Convolutional Neural Networks, CNN)和循环神经网络^[4](Recurrent Neural Networks, RNN)为深度学习网络中的两大主流网络。在情感分析方面, Yang 等^[5]利用网络术语和维基中文数据集扩展原始词汇,并根据句子的长度对卷积神经网络(CNN)的池化层进行优化,实现情感分析任务。Niu 等^[6]提出了基于 Word2Vec 和针对长短期记忆网络改进的循环神经网络,并将其应用在微博文本中,实验证明该模型优于 SVM 模型、RNN 模型、CNN 模型。为了进一步提升神经网络模型在情感分析方面的性能,并更好地关注文本中对情感类别贡献较为关键的特征,研究人员将 Attention 机制^[7]引入神经网络。Wu 等^[8]利用 Self-Atten-

到稿日期:2020-10-13 返修日期:2021-03-31

基金项目:国家自然科学基金(61532008);国家重点研发计划(2017YFC0909502)

This work was supported by the National Natural Science Foundation of China (61532008) and National Key R&D Program of China (2017YFC0909502).

通信作者:张亚文(1754107561@qq.com)

tion 和 Bi-LSTM 提出基于字向量的中文文本情感分析,在微博数据集上的分类效果比单独使用 LSTM、CNN 等神经网络提高了 1.15%。Zhai 等^[9]将注意力机制和 Bi-LSTM 相结合,用于双语的文本情感分析,并取得了很好的成绩。但这些模型都是将注意力机制和神经网络进行拼接,没有对网络结构进行改进。为了在联系上下文、加强文本中重要词汇对情感倾向的影响方面取得更好的效果,本文参考文献[10]中注意力机制和 GRU 相结合的方式,提出了具有注意力更新门的 BiGRU 模型,将注意力机制引入门控单元的更新门,从而提高模型提取重要信息的能力,并将新的模型与强制向前注意力机制结合,进一步提升模型的性能。

2 相关研究

2.1 GloVe 词向量

使用神经网络学习文本特征首先需要将文本中的单词转换为向量,传统的向量化方法为 one-hot 编码^[11],但该方法忽略了词与词之间的关系,容易造成维度灾难。2013 年, Mikolov 等^[12]提出 Word2Vec 模型来建立神经网络词嵌入表示,此模型能有效降低单词维度,使其值为 50~100。Word2Vec 模型主要基于 CBOW(Continuous Bag of Word)和 skip-gram 模型两种方式实现。虽然 skip-gram 模型在计算近义词方面比较出色,但它只是在局部上下文窗口训练模型,并且很少使用语料中的一些统计信息,因此 Pennington 等^[13]又提出了一个新型模型 GloVe。与 Word2Vec 相比, GloVe 利用了全局信息,使其在训练时收敛更快,训练周期较 Word2Vec 更短,且效果更好。

2.2 BiGRU 结构模型

为了弥补 LSTM 训练时间长、参数较多、内部计算复杂的缺点, Chung 等^[14]于 2014 年提出了一种改进模型,即 GRU(Gated Recurrent Unit)模型。GRU 模型将 LSTM 的单元状态和隐层状态进行合并,并将遗忘门和输入门合成一个单一的更新门,简化了其结构,很大程度上提高了训练效率。GRU 模型如图 1 所示,以 r_t 表示重置门, z_t 表示更新门,将单元状态与输出合并为一个状态 h_t 。

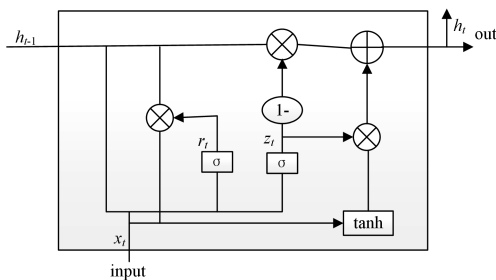


图 1 GRU 结构模型图

Fig. 1 GRU structure model diagram

GRU 的向前计算公式为:

$$z_t = \sigma(w_z \cdot [h_{t-1}, x_t]) \quad (1)$$

$$r_t = \sigma(w_r \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(w_h \cdot [r_t * h_{t-1}, x_t]) \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (4)$$

其中, σ 为 sigmoid 激活函数; x_t 为第 t 步的输入单元,表示第 t 个单词的词向量; h_t 为第 t 步的隐藏状态; h_{t-1} 为 $t-1$ 时刻的隐藏状态; r_t 表示重置门; z_t 表示更新门; \tilde{h}_t 为 t 时刻的候选记忆单元。

在文本情感分析中,我们希望当前时刻的输出能与上下文的状态都产生联系,但是单向的 GRU 只能捕捉上文信息,针对此缺陷,本文使用双向的 GRU 模型,将正向隐含层和反向隐含层相结合,从而实现上下文之间的联系。BiGRU 模型的结构如图 2 所示。BiGRU 当前的隐层状态由当前的输入 x_t 、 $t-1$ 时刻向前的隐层状态的输入、反向的隐层状态的输出这 3 个部分共同决定。

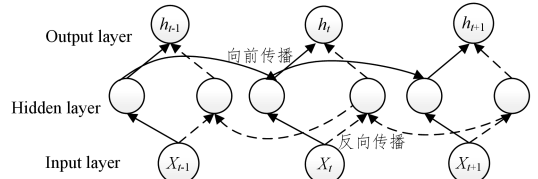


图 2 BiGRU 结构模型图

Fig. 2 BiGRU structure model diagram

3 融合注意力机制和 BiGRU 的 FFA-BiAGRU 文本情感分析模型

为了提高文本情感分析的准确性,本文结合 BiGRU 模型和 Attention 机制,设计了一个新的情感分析模型 FFA-BiAGRU,该情感分析模型如图 3 所示,主要包括文本预处理、向量化和分类器 3 个部分。

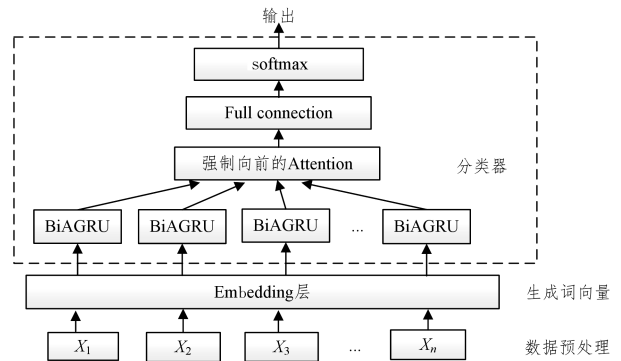


图 3 FFA-BiAGRU 情感分析模型

Fig. 3 FFA-BiAGRU sentiment analysis model

在文本预处理阶段,首先对文本数据进行去停用词、去除换行符等清洗工作,统一英文数据集中英文字母的大小写,并将数据序列化;然后将处理好的数据通过 GloVe 向量化,将文本数据转换成向量;最后将词向量放入分类器中进行处理。在该部分使用改进的 BiAGRU 模型和强制向前的注意力机制, BiAGRU 模型对数据进行学习,提取出重要特征;强制向前的 Attention 整合文本中的重要信息,最后通过全连接层得到分类结果。

3.1 FFA-BiAGRU 模型

FFA-BiAGRU 模型的实现过程如图 4 所示,通过 GloVe 词向量得到的词向量为 BiAGRU 部分的输入,某一时刻的词

向量 h_t 通过注意力机制得到注意力得分 a_t , a_t 作用于 BiGRU 中的隐层,影响更新门的数值大小。最后,将得到的特征向量输入到强制向前的 Attention 中,将重要信息进行整合,得到向量 C 。

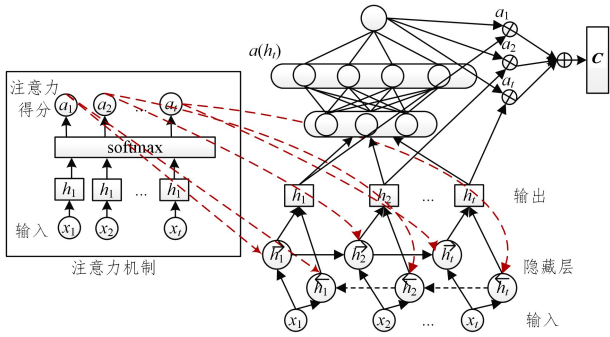


图4 FFA-BiAGRU 的模型结构

Fig. 4 FFA-BiAGRU model structure

3.2 Embedding 层

Embedding 层即词嵌入层。在对文本进行分析前,需要将文本中的文字转换为神经网络可以处理的词向量。本文使用 GloVe 来实现词语向量化,通过该模型将词语转换为 100 维的词向量。GloVe 的本质是对共现矩阵进行降维操作。模型的公式如下:

$$J = \sum_{i,j} f(X_{i,j}) (v_i^T v_j + b_i + b_j - \log(X_{i,j}))^2 \quad (5)$$

其中, v_i 和 v_j 是单词 i 和单词 j 的词向量, b_i 和 b_j 是两个偏差项, f 是权重函数, N 是词汇表大小, X 为共现矩阵, X_{ij} 为整个语料库中单词 i 和单词 j 共同出现在一个窗口中的次数。

具体的权重函数如下:

$$f(x) = \begin{cases} (x/x_{\max})^{0.75}, & \text{if } x < x_{\max} \\ 1, & \text{if } x > x_{\max} \end{cases} \quad (6)$$

GloVe 模型可通过式(6)直接使用自身语料库进行计算。

3.3 BiAGRU 层

3.3.1 Attention 机制

注意力机制是一种类似人脑的注意力分配机制,它对重要的区域投入更多的资源,以获取更多的细节,对无用的信息则进行抑制。该部分的 Attention 机制的实现公式如下:

$$u_i = \tanh(W_w h_i + b_w) \quad (7)$$

$$a_i = \text{softmax}(u_i) \quad (8)$$

其中, h_i 为第 i 个窗口产生的特征, W_w 和 b_w 为 Attention 的权重和偏置向, a_i 为注意力分数。因为该部分仅需注意力分数对更新门进行优化,所以不需要得出各词的注意力权重。

3.3.2 注意力门控单元 BiAGRU

为了弥补 LSTM 训练时间长、参数较多、结构复杂的缺点,提出了 GRU 模型。GRU 模型主要根据更新门和重置门两个门控系统对输入信息进行筛选,更新门的值越大说明前一时刻的隐层输出对当前隐层的影响越大,更新门的值越小说明忽略得越多。但传统的更新门主要由前一时刻的历史信息和新接收的信息决定,当输入序列较长时,其并不能很好地提取文本中的重要信息。为了进一步提高对输入信息的敏感

性和 GRU 模型提取特征的能力,有学者提出将 Attention 机制和 GRU 相结合来提取特征向量。文献[10]对 Attention 机制和 GRU 不同的结合方式对兴趣演化的影响进行实验,得出将 Attention 机制嵌入 GRU 中更为有效,本文也将采取嵌入的方式对情感分类模型进行改进。

改进后的 AGRU 模型结构如图 5 所示。改进后的更新门不仅与前一刻的信息和新输入信息有关,还与当前信息的注意力得分有关。信息的影响力越大,注意力得分越高,更新门的值就越大,信息被保留,反之将被丢弃,从而更好地提取文本中的重要信息。

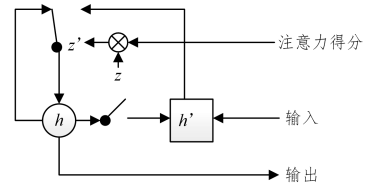


图5 AGRU 结构模型图

Fig. 5 AGRU structure model diagram

AGRU 模型的实现公式如下:

$$z_t = \sigma(w_z \cdot [h_{t-1}, x_t]) \quad (9)$$

$$z'_t = a_t * z_t \quad (10)$$

$$r_t = \sigma(w_r \cdot [h_{t-1}, x_t]) \quad (11)$$

$$\tilde{h}_t = \tanh(w_h \cdot [r_t * h_{t-1}, x_t]) \quad (12)$$

$$h'_t = (1 - z'_t) * h_{t-1} + z'_t * \tilde{h}_t \quad (13)$$

其中, z_t 是 AGRU 的原始更新门, z'_t 是 AGRU 模型的注意更新门, h'_t 和 h_{t-1} 是 AGRU 的隐藏状态。

为了进一步提高模型的精确度,本文将 AGRU 模型双向化后得到 BiAGRU 模型。BiAGRU 模型结构与 BiGRU 模型结构相似,如图 2 所示。给定单词词向量为 $w_{i,t} \in [1, L]$, L 为文本长度, W_e 为 BiAGRU 的权重矩阵,则文本单词向量化表示为:

$$x_t = W_e w_{i,t}, t \in [1, L] \quad (14)$$

BiAGRU 的计算公式为:

$$\vec{h}_t = \text{AGRU}(x_t), t \in [1, L] \quad (15)$$

$$\overleftarrow{h}_t = \text{AGRU}(x_t), t \in [L, 1] \quad (16)$$

其中, \vec{h} 表示向前传播时词语的隐状态, \overleftarrow{h} 表示反向传播时词语的隐状态。将其拼接为 $h_t = [\vec{h}, \overleftarrow{h}]$, 得到词向量双向语义信息。

3.4 强制向前的 Attention 层

与前文使用的 Attention 不同,此处使用的是强制向前的注意力(Feed-Forward Attention)机制。强制向前的 Attention 机制要求在生成目标单词时,如果某个输入单词已经和输出单词对齐,则后面基本不再考虑。该注意力机制对向量的加权求和方式进行了创新,实现公式如下:

$$e_t = a(h_t) \quad (17)$$

$$a_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)} \quad (18)$$

$$c = \sum_{t=1}^T a_t \mathbf{h}_t \quad (19)$$

其中, \mathbf{h}_t 为 t 时刻的隐藏状态; a 为学习函数, 用于计算 \mathbf{h}_t 的重要性; T 为输入序列时长的总步数。

4 实验与分析

4.1 数据预处理

本文使用的是 IMDB 电影评论数据集和来自今日头条客户端的文本分类数据集。数据集 1 共有 25 000 条数据, 被标记为积极和消极两类; 数据集 2 共有 382 688 条数据, 分为 15 个类型。首先, 对数据集 1 进行预处理, 过滤掉非 ASCII 字符、清洗换行符以及将大写字母转换为小写, 并使用 GloVe 初始化评论文本的词嵌入信息; 然后将数据集 1 按 8:2 的比例随机分为训练集和测试集, 分别用于模型的训练和性能评估, 其中训练集有 20 000 条数据, 测试集有 5 000 条数据。

4.2 评价指标

本文采用准确率(accuracy)、召回率(recall)和 F1 作为评价指标, 计算式如下:

$$accuracy = \frac{TP + TF}{TP + TF + FP + FN} \quad (20)$$

$$recall = \frac{TP}{TP + FN} \quad (21)$$

$$precision = \frac{TP}{TP + FP} \quad (22)$$

$$F1 = \frac{2 * recall * precision}{precision + recall} \quad (23)$$

其中, TP 表示预测为正样本且分类正确的样本数, TF 表示预测为负样本且分类正确的样本数, FP 表示实际为负但分类错误的样本数, FN 表示实际为正但分类错误的样本数。

4.3 对比实验和参数设置

本实验将 FFA-BiAGRU 模型与以下 4 种模型进行对比。

(1) 文献[15]提出的 GloVe-GRU 模型, 该模型将 GloVe 词向量和 GRU 文本分类器相结合。

(2) 文献[16]提出的 BiGRU-Attention 神经网络模型。

(3) 文献[17]提出的 BiLSTM-Attentions 模型。该模型将 BiLSTM 与两层 Attention 相结合。

(4) BiLSTM-Attention 模型与(3)中的模型是同类型模型, 不同点是该模型只与一层 Attention 相结合。

所有实验均分为数据预处理、特征表示、文本分类 3 部分, 且预处理过程、超参数保持一致。

本实验模型的参数设置如表 1 所列。

表 1 模型参数设置表

Table 1 Model parameter setting table

参数名称	参数值
词向量维度	100
学习率	0.01
损失函数	binary_crossentropy
Batch size	100
Dropout	0.2
optimizer	SGD

4.4 IMDB 数据集的实验结果分析

各模型在 IMDB 数据集上的分析结果如表 2 所列, 各

别的分类结果如图 6 所示。

表 2 各模型的情感分析结果

Table 2 Sentiment analysis results of each model

Models	accuracy	recall	F1
BiLSTM-Attentions	0.8836	0.8804	0.8864
GloVe-GRU	0.8496	0.8389	0.8481
BiGRU-Attention	0.8923	0.8921	0.8924
BiLSTM-Attention	0.8648	0.8690	0.8654
FFA-BiAGRU	0.9012	0.8988	0.9011

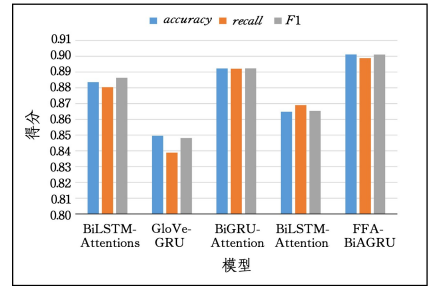


图 6 各模型的结果对比

Fig. 6 Comparison of various model results

由图 6 可知, 本文提出的实现模型在准确率、召回率和 F1 值上都优于其他模型。分析表 2 和图 6 可知, BiLSTM 模型优于 GRU 模型, 这说明双向模型优于单向模型; BiLSTM 与双层 Attention 结合的模型比 BiLSTM-Attention 模型的效果好, 这两种模型都使用了 BiLSTM 和 Attention 来提取文本信息, 不同点在于 Attention 的层数不同, 这说明 Attention 层是有效的; BiGRU-Attention 模型和 BiLSTM-Attention 模型相比, 两者都有 Attention 层, 模型结构基本相同, 但 BiGRU-Attention 的得分更高, 说明在小数据集上, BiGRU 在提取文本信息方面优于 BiLSTM。在此基础上, 本文对 BiGRU 进行了进一步的改进, 结果表明, 本文模型在各方面均表现优异。

4.5 文本分类实验结果分析

为了进一步验证模型的有效性, 本文在文本分类数据集上再次进行实验。本实验使用的数据集来自今日头条客户端, 共包含 382 688 条数据, 包括民生故事、文化、娱乐、体育、财经、房产、汽车、教育、科技、军事、旅游、国际、证券股票、农业和电竞游戏 15 种类型。为了更好地进行文本分类, 本文对原始数据集进行处理, 只保留文章类型和标题, 以“_!”进行分割。对数据进行去停用词、去换行符等预处理后, 按 8:1:1 的比例随机划分训练集、测试集和验证集。数据集结构如表 3 所列。

表 3 数据集结构表

Table 3 Data set structure table

数据集	(单位: 条)		
	训练集	测试集	验证集
中文文本分类数据集	306150	38269	38269

本节在 CNN 模型、BiLSTM-Attentions 模型、BiLSTM-Attentions 模型、BiGRU-Attention 模型以及本文提出的改进模型上对数据集中的新闻文本进行实验, 结果如图 7 所示。

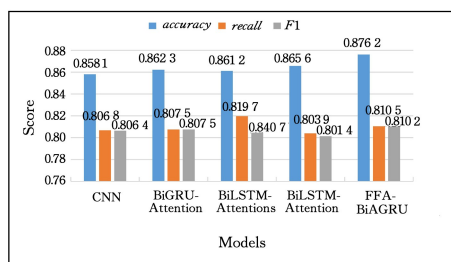


图7 各模型的结果对比

Fig.7 Comparison of various model results

由图7可知,几种模型在准确率、召回率、F1值方面差距不大,这说明当数据集较大时GRU模型与LSTM模型的功能接近。添加了Attention层的模型明显优于CNN模型,这说明提高网络的复杂度是有效的。虽然BiLSTM-Attention模型的准确率较高,但其召回率和F1值较低,综合来看BiGRU-Attention模型的效果更好。实验结果表明,本文提出的FFA-BiAGRU模型能更好地联系上下文,同时结合强制向前Attention机制提取深层特征的模型在准确率和F1值方面都优于其他模型。

结束语 传统的情感分析模型没有充分考虑上下文之间的关系和文本中重要词汇对情感分析的影响,大多数模型为神经网络模型与Attention机制的堆叠。针对这些缺陷,本文提出具有注意力更新门的BiGRU网络模型,利用注意力得分来控制更新门,对模型进行优化,并与强制向前Attention机制相结合,进一步提高模型的准确率。实验结果表明,本文模型具有较好的有效性。

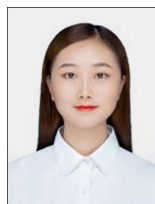
在未来的工作中,会进一步考虑不同的Attention机制与GRU融合的情况,并对损失函数进行优化,尝试对模型进行改进,在更多的领域中证明模型的有效性。

参考文献

- [1] JHA V, SAVITHA R, SHENOY P D, et al. A novel sentiment aware dictionary for multi-domain sentiment classification[J]. Computers & Electrical Engineering, 2017, 69: 585-597.
- [2] XU J Z, ZHU J, ZHAO R, et al. Sentiment Analysis of Aerospace Microblog Using SVM[J]. Journal of Information Security Research, 2017, 3(12): 1129-1133.
- [3] LI P, ZHAO W D. Image fire detection algorithms based on convolutional neural networks[J]. Case Studies in Thermal Engineering, 2020, 19: 100625.
- [4] TEHSEEN Z, ASSAD A, USMAN H, et al. Learning deep hierarchical and temporal recurrent neural networks with residual learning[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(2): 873-882.
- [5] YANG X Y L, XU S J, WU H, et al. Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network[J]. Procedia Computer Science, 2019, 147: 361-368.
- [6] NIU C M, ZHANG G H, LI Z H. Chinese Weibo Sentiment Analysis Based on Deep Neural Network[J]. Computer Systems & Applications, 2018, 27(11): 205-210.
- [7] LIAO W X, ZENG B, XU Y Y. Natural language processing model based on one-dimensional dilated Convolution and attention mechanism[J/OL]. Computer Engineering and Applications, 1-8. [2020-05-20]. <http://kns.cnki.net/kcms/detail/11.2127.TP.20200422.0950.002.html>.
- [8] WU X H, CHEN L, WEI T T, et al. Sentiment Analysis of Chinese Short Text Based on Self-Attention and Bi-LSTM[J]. Journal of Chinese Information Processing, 2019, 33(6): 100-107.
- [9] ZHAI S P, YANG Y Y, QIU C, et al. Bilingual text sentiment analysis based on attention mechanism Bi-LSTM[J]. Computer Applications and Software, 2019, 36(12): 251-255.
- [10] ZHOU G, MOU N, FAN Y, et al. Deep Interest Evolution Network for Click-Through Rate Prediction[J]. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 2018, 33(1): 5941-5948.
- [11] FU Y X, LU T L, MA Z L. CNN malicious code detection technology based on one-hot[J]. Computer Applications and Software, 2020, 37(1): 304-308.
- [12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Ence, 2013. arXiv:1301.3781.
- [13] PENNINGTON J, SOCHER R, MANNING C. Glove: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.
- [14] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J/OL]. Eprint Arxiv, 2014. <https://arxiv.org/pdf/1412.3555.pdf>.
- [15] FANG J K, CHEN P H, LIAO W X. Text classification model based on GloVe and GRU[J/OL]. Computer Engineering and Applications. [2020-08-13]. <http://kns.cnki.net/kcms/detail/11.2127.tp.20200331.1742.006.html>.
- [16] WANG W, SUN Y X, QI Q J, et al. Text sentiment classification model based on BiGRU-Attention neural network[J]. Application Research of Computers, 2019, 36(12): 3558-3564.
- [17] KUMAR A, KAWAHARA D, KUROHASHI S, et al. Knowledge-enriched two-layered attention for sentiment analysis[C]// North American Chapter of the Association for Computational Linguistics. 2018: 253-258.



YANG Qing, born in 1965, master, associate professor, is a member of China Computer Federation. Her main research interests include data mining and computer application technology.



ZHANG Ya-wen, born in 1995, master. Her main research interests include data mining and computer application technology.