

## 基于 DSR 和 BGRU 模型的聊天文本证据分类方法

张宇<sup>1</sup>, 李炳龙<sup>1</sup>, 李学娟<sup>2</sup>, 张和禹<sup>1</sup>

(1. 信息工程大学, 河南 郑州 450001; 2. 河南理工大学, 河南 焦作 454003)

**摘要:** 即时通信等社交软件产生的聊天文本内容证据数据量大且聊天内容含有“黑话”等复杂语义, 数字取证时无法快速识别和提取与犯罪事件有关的聊天文本证据。为此, 基于 DSR (dynamic semantic representation) 模型和 BGRU (bidirectional gated recurrent unit) 模型提出一个聊天文本证据分类模型 (DSR-BGRU)。通过预处理手段处理聊天文本数据, 使其保存犯罪领域特征。设计并实现了基于 DSR 模型的聊天文本证据语义特征表示方法, 从语义层面对聊天文本进行特征表示, 通过聚类算法筛选出语义词, 并通过单词属性与语义词的加权组合对非语义词词向量进行特征表示, 且将语义词用于对新单词进行稀疏表示。利用 Keras 框架构建了包含 DSR 模型输入层、BGRU 模型隐藏层和 softmax 分类层的多层聊天文本特征提取与分类模型, 该模型使用 DSR 模型进行词的向量表示组成的文本矩阵作为输入向量, 从语义层面对聊天文本进行特征表示, 基于 BGRU 模型的多层隐藏层对使用这些词向量组成的文本提取上下文特征, 从而能够更好地准确理解聊天文本的语义信息, 并利用 softmax 分类层实现聊天文本证据识别与提取目标。实验结果表明, 基于 DSR-BGRU 的聊天文本证据分类模型能够更加准确地完成聊天记录证据的识别和提取任务, 该模型能够有效地提取出聊天信息中的犯罪文本信息, 取得有效的证据, 并取得了 92.06% 的准确率, F1 值为 91.00%。高于其他用于文本分类的模型与方法。

**关键词:** 文本语义表示; 一词多义; 文本分类; 数字取证

**中图分类号:** TP391

**文献标志码:** A

**DOI:** 10.11959/j.issn.2096-109x.2022007

## Evidence classification method of chat text based on DSR and BGRU model

ZHANG Yu<sup>1</sup>, LI Binglong<sup>1</sup>, LI Xuejuan<sup>2</sup>, ZHANG Heyu<sup>1</sup>

1. Information Engineering University, Zhengzhou 450001, China

2. Henan Polytechnic University, Jiaozuo 454003, China

**Abstract:** It is always unlikely to efficiently identify and extract chat text evidence related to criminal events, due to

收稿日期: 2021-03-08; 修回日期: 2021-06-10

通信作者: 李炳龙, lbl2017@163.com

基金项目: 国家自然科学基金 (60903220)

**Foundation Item:** The National Natural Science Foundation of China(60903220)

**引用格式:** 张宇, 李炳龙, 李学娟, 等. 基于 DSR 和 BGRU 模型的聊天文本证据分类方法[J]. 网络与信息安全学报, 2022, 8(2): 150-159.

**Citation Format:** ZHANG Y, LI B L, LI X J, et al. Evidence classification method of chat text based on DSR and BGRU model[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 150-159.

the complex semantics such as “slang” in the chat content and the huge amount of chat text data generated by social software such as instant messaging. Based on this motivation, a chat text evidence classification model (DSR-BGRU) based on the DSR (dynamic semantic representation) model and the BGRU (bidirectional gated recurrent unit) model was proposed. The chat text data was pre-processed to preserve the characteristics of the criminal field. Then a multi-layer chat text feature extraction and classification model using the Keras framework was proposed. With the text matrix composed of vector representation of words in the DSR model as the input vector, the input layer of the DSR model featured the chat text from the semantic level. Then the hidden layer of the BGRU model extracted the context characteristics of the text composed of the word vectors. The softmax classification layer recognized and extracted the chat text evidence. The experimental results show that the proposed DSR-BGRU can more accurately identify and extract chat records compared with other models and methods for text classification, and it can also effectively extract the criminal text information from the chat information with the accuracy rate 92.06% and the  $F1$  score 91.00%.

**Keywords:** text semantic representation, polysemy, text classification, digital forensics

## 0 引言

随着移动互联网的快速发展,智能手机的使用频率也在快速增长。但智能手机在给人们带来便利的同时,也给犯罪带来了便利。利用智能手机犯罪的记录在不断增长。以往的手机取证主要关注手机保存的地理信息数据、短信、通话记录等。但随着社交软件的发展,社交软件所保留的聊天记录已经成为新的取证来源。根据 CNNIC 发布的调查报告显示<sup>[1]</sup>,截至 2021 年 1 月,我国社交软件的使用人数已经达到 96 012 万。正是由于社交软件的流行,社交软件上的聊天记录具有重要的取证意义。

取证人员通过对聊天记录进行分析来找到与案情相关的数据,但问题在于聊天记录大量繁杂,而取证人员采用的方式是手动搜索或者通过设置过滤器的方法,这样的方法低效而且很可能会遗漏关键证据。此外,近年来,犯罪分子的反取证意识不断提高,在使用社交软件时往往采用“黑话”来传递犯罪信息。例如,在非法借贷中“高炮”指一种超高息的短期借款;电信诈骗中“金主”指实施诈骗者,“水房”指洗钱机构;毒品交易中“猪肉”“苹果”等指冰毒,“溜冰”指吸食冰毒。“黑话”影响了取证的效率。因此需要一种文本分类方法来从犯罪领域的聊天记录中筛选出关键数据。

将文本分类技术应用到聊天记录中仍有许多问题需要解决。

第一,犯罪领域的聊天文本具有鲜明的犯罪特征,常用的聊天记录预处理方法会丢失这些特征。

第二,随着时间的移动,“黑话”不断出现并更新。出现新单词会降低分类器的准确性,某些语义信息会在文本量化期间被忽略。

第三,“黑话”,如“猪肉”“苹果”等涉毒词,往往与涉及的领域无关,这需要结合上下文信息才能被认为与犯罪行为有关。这种“一词多义”问题影响分类器的效果。

为了将文本分类技术应用到聊天记录来筛选与犯罪有关的文本,本文提出了一种聊天记录犯罪证据分类方法。该方法通过更新文本特征表示和文本特征提取技术来提高分类性能。

针对犯罪相关的文本分类主要应用于社交平台所产生的聊天记录。文献[2]提出 ARPR (AHP-relation-PageRank) 算法。该算法采用 TF-IDF (term frequency-inverse document frequency) 方法对涉毒关键词进行提取并对其词向量加权,然后以好友关系为链接建立关系网络作为 PageRank 的入度与出度来计算相对应的 PageRank 权值。ARPR 算法结合 TF-IDF 方法和 PageRank 方法对聊天文本进行处理,使其能根据禁毒领域的业务需求,筛选出聊天文本中涉及的重点嫌疑人员。

与其他文本分类任务相比,与犯罪相关的文本特点是存在大量的“黑话”(一词多义)现象。有不少研究者将深度学习用于解决文本分类中的

一词多义问题<sup>[3-4]</sup>。文献[5]引入 RNN (recurrent neural network) 抽取词语特征, 其中词语的特征不仅与当前词有关还与它的上文有关。这种做法使得词语的特征表示会随着上文的而变化而变化, 具有语义的特性。但 RNN 循环机制过于简单, 在梯度反向传播时会产生梯度爆炸或梯度消失等错误状况, 从而造成训练过程停滞。LSTM (long short-term memory) 和 GRU (gated recurrent unit) 在传统 RNN 的基础上引入“门机制”, 在一定程度上避免了 RNN 的失误。以此为基础, 大部分模型对 GRU 和 LSTM 模型实施改良。文献[6]为了增强模型中上下文信息的处理效果, 在 LSTM 模型的基础上新增了外部记忆单元。但外部记忆单元参数量大, 容易影响准确度。Tang 等<sup>[7]</sup>提出 CLSTM 模型来解决一词多义问题, 该模型在 LSTM 的基础上, 引入注意力机制, 取得了不错的效果。文献[8]在文献[9]的基础上进行改进, 引入注意力机制来聚焦文本中的重要信息。文献[10]与文献[8]类似, 区别在于, 其引入 BGRU 来代替 LSTM, 相比文献[8], 文献[10]提出的模型在时间性能上有较大提升。

以上的研究将 RNN、CNN 这样的模型结构串联起来用于文本分类, 但这种层叠串联的模型结构会随着层次的加深丢失特征信息。针对这一问题, 文献[11]提出一种双通道深度学习模型, 该模型并行采用 CNN 和 LSTM 分别从字、词角度对不同嵌入方式下的文本进行特征提取, 并将双通道提取出的特征进行融合, 从而进行分类。文献[12]将 CNN 和 LSTM 并联起来, 并在模型的后端综合各通道提取的特征进行分类。

总而言之, 本文的研究技术流程如图 1 所示。本文的主要贡献如下。

1) 提出犯罪领域的聊天文本预处理的手段, 使其保存犯罪领域特征。

2) 提出动态语义表示模型 (DSR) 对词进行特征表示。该模型首先在预训练好的词向量上使用聚类算法选择语义词。然后采用稀疏约束训练语义词对新单词进行稀疏表示。采用 DSR 模型能有效降低新单词对分类器准确性的影响。

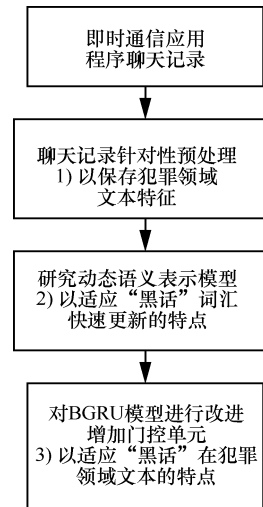


图 1 研究技术流程

Figure 1 Research technology process

3) 采用 BGRU 模型<sup>[13]</sup>从词向量组成的文本中进行上下文特征提取, 并对其改进使其能适应“黑话特点”从而能根据上下文变化来解决“黑话”一词多义问题。

## 1 模型设计

为从犯罪领域的聊天记录中筛选出关键数据, 本文提出结合 DSR 模型和 BGRU 模型的聊天记录文本分类方法。其中, DSR 模型用于捕获语义层次的信息, BGRU 模型提取词向量的文本特征并用于分类。首先对聊天记录进行分类, 分为训练集和测试集, 在预处理后, 先训练训练集, 而后对结果进行测试, 通过反馈的正确率验证模型的准确率, 来看模型是否有效。DSR-BGRU 模型结构如图 2 所示, 主要分为 4 个模块: 聊天记录预处理; DSR 模型进行文本特征表示; BGRU 提取文本特征; 实施文本分类。

### 1.1 聊天记录预处理

1) 在聊天记录中, 除了文本数据外, 还存在符号、表情、图片、语气词等非文本数据, 这些非文本数据将会变成乱码而影响数据处理的结果, 因此应先去除这类数据。但犯罪领域的聊天记录中的图片内容往往是展示用于犯罪的道具, 或者触犯法律的物品。这类图片具有较高的取证价值以及鲜明的犯罪领域特征。DSR 模型采用识图工具识别图片内容, 并将其输出为文本替代聊天记录中的图片。

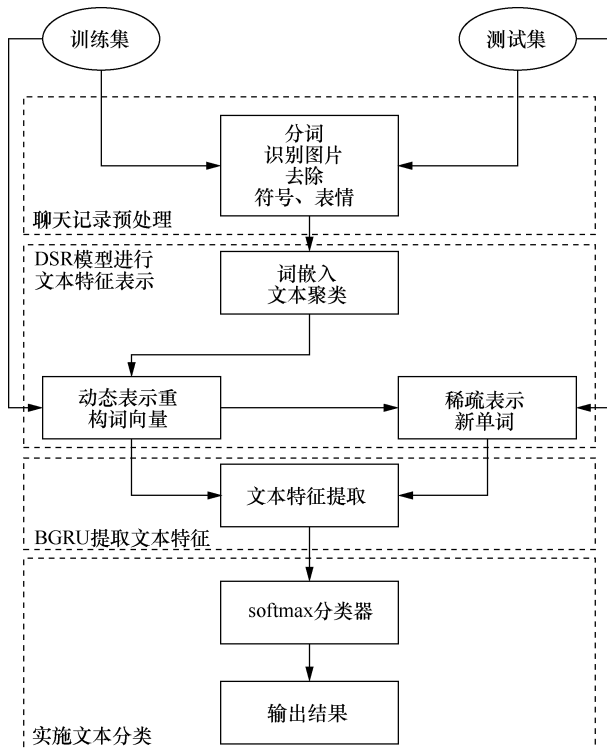


图 2 DSR-BGRU 模型结构

Figure 2 The structure of DSR-BGRU model

2) 犯罪分子在聊天记录中会用“黑话”来掩盖自己的犯罪意图，这类词往往具有名词的特性。DSR 模型对聊天记录分词切割后标注词性用于后续的加权，标注结果如表 1 所示，词性对应含义如表 2 所示。

表 1 分词及词性标注结果

Table 1 Word segmentation and speech tagging results

| 原始聊天记录     | 切词及词性标注后文本                    |
|------------|-------------------------------|
| 有猪肉么       | 有/v 猪肉/n 么/o                  |
| 有信么，我这有洗信人 | 有/v 信/n 么/o，我/o 这/o 有/v 洗信人/n |
| 四号到货了，要溜么  | 四号/n 到/v 货/n 了/o，要/v 溜/v 么/o  |

表 2 词性对应含义

Table 2 Corresponding meaning of word

| 词性 | 对应含义 |
|----|------|
| n  | 名词   |
| v  | 动词   |
| o  | 其他   |

## 1.2 文本特征表示

本文提出 DSR 模型对文本进行特征表示。DSR 模型通过聚类算法筛选出语义词，并通过单词属性与语义词的加权组合对非语义词词向量进行特征

表示，且将语义词用于对新单词进行稀疏表示。

### 1.2.1 语义词筛选与重构

语义词的筛选基于已经预训练好的词向量。采用预训练的词向量，模型将每个单词的嵌入向量  $w$  组成矩阵。然后对聊天信息  $T$  进行聚类分析，将具有相似语义的单词划分为同一类。在每个聚类过程中，提取聚类中心作为文本的语义词。

本文使用 AP 聚类算法计算文本的聚类中心并作为语义词，假设文本的嵌入向量集为  $\{w_1, w_2, \dots, w_n\}$ ，将  $S$  定义为词向量之间的相似矩阵， $s(i, j)$  表示  $w_i$  和  $w_j$  两个词向量之间的相同程度，值为两个词向量之间的欧式距离。信息矩阵用  $A$  来表示，而吸引信息矩阵用  $R$  来表示， $R$  中  $r(i, j)$  描述  $w_j$  适合作为  $w_i$  的聚类中心的适用性， $A$  中  $a(i, j)$  描述  $w_i$  选择  $w_j$  作为其聚类中心的适用性，如式(1)~式(4)所示。

$$r_{t+1}(i, k) = \begin{cases} s(i, k) - \max_{j \neq k} \left\{ a_t(i, j) + r_t(i, j) \right\}, & i \neq k \\ s(i, k) - \max_{j \neq k} \{ s(i, j) \}, & i = k \end{cases} \quad (1)$$

$$a_{t+1}(i, k) = \begin{cases} 0, r_{t+1}(k, k) + \min \left\{ \sum_{j \neq \{i, k\}} \max \{ 0, r_{t+1}(i', k) \} \right\}, & i \neq k \\ \sum_{j \neq k} \max \{ r_{t+1}(j, k), 0 \}, & i = k \end{cases} \quad (2)$$

$$r_{t+1}(i, k) = (1 - \lambda)r_{t+1}(i, k) + \lambda r_t(i, k) \quad (3)$$

$$a_{t+1}(i, k) = (1 - \lambda)a_{t+1}(i, k) + \lambda a_t(i, k) \quad (4)$$

若在多次迭代后聚类中心保持不变或迭代次数超过了设置的迭代次数，将停止计算，并将计算出的聚类中心作为语义词，最终取  $A + R$  最大的  $w_k$  作为聚类中心。

基于上述方法，DSR 模型可以将文本中相似的词语替换为相应的语义词并且对训练集中的每个文本都生成了语义词集合  $SW$ 。该集合包含  $m$  个语义词  $\{sw_1, sw_2, \dots, sw_m\}$ 。训练集中文本的语义词集合会组成词库，用于对测试集中的新单词进行表示，从而进一步提高了模型的自适应能力。

模型结合词的各项特征以及与词相近的语义词来表示训练集中的非语义词词向量。采用式(5)

对语义词集合的词进行特征融合。

$$V(w(i)) = \text{MTF}(w_i) \quad (5)$$

其中,  $\text{MTF}()$  是特征融合函数。不同于仅使用统计特征来表示所有单词的传统单词袋 (BoW), DSR 模型结合单词频率、单词词性、单词位置这 3 个特征来重构词向量。

单词频率是单词在文本中出现的次数。单词频率越高, 单词越重要。它是统计功能中常用的词属性之一。模型运用函数来计算词频因子  $\text{fre}_i$ , 如下所示。

$$\text{fre}_i = \frac{n}{n+1} \quad (6)$$

$n$  是单词  $w_i$  在文章中的出现次数。非线性函数有两个优点: 一是词频因子与词频成正比; 二是词频因子与词频成正比。当词频增加到一定程度时, 词频因子的值会减小, 符合语言现实。

词性因子是词性的量化。对文本语义有较大影响的语义词大多数是名词。与名词相比, 形容词和动词对句子不会产生太大的影响。在对文本的分类中, 词性的不同产生的影响也不一样, 因此根据词性将词分为 3 类:

$$\text{pos}_i = \begin{cases} 1, & w_i \text{ 为名词} \\ 0.8, & w_i \text{ 为动词或形容词} \\ 0.6, & \text{其他} \end{cases} \quad (7)$$

单词在文本中的位置在判断其重要性方面也具有价值。不同的单词出现在文本的不同位置, 对语义的影响也不同。根据如下公式定义位置的影响力:

$$\text{loc}_i = \frac{\text{last}_i - \text{first}_i + 1}{\text{sum}_i} \quad (8)$$

$\text{first}_i$  是  $w_i$  首先出现的位置,  $\text{last}_i$  是  $w_i$  最后出现的位置,  $\text{sum}_i$  是文本中单词的总数。

这里, 构造融合特征来表示语义词, 计算如下。

$$\text{MTF}(w_i) = (\alpha_1 \text{fre}_i + \alpha_2 \text{pos}_i + \alpha_3 \text{loc}_i) s(w_i) \quad (9)$$

其中,  $\text{fre}_i$  是单词频率因子,  $\text{pos}_i$  是词性因子,  $\text{loc}_i$  是单词位置因子,  $\alpha_1$ 、 $\alpha_2$  和  $\alpha_3$  是特征因子的权重,  $S(w_i)$  表示 SW 中与  $w_i$  相近的语义词。因此, 根据式(5), 可以使用  $\{V(w(1)), V(w(2)), \dots,$

$V(w(i))\}_{i=1}^n$  来表示具有  $i$  个词的文本  $T_j$ 。

### 1.2.2 稀疏表示

由于训练集和测试集是随机分配的, 并且词库中的语义词由训练集中的词组成, 而从测试集提取的词可能不会出现在词库中。如果从测试文本中提取的词未出现在词库中, 则将使用其他语义词来稀疏地表示这些词。目标函数如下。

$$\underset{\alpha}{\text{argmin}} \|y - \alpha X\|_2^2, \text{ s.t. } \|\alpha\|_1 < \varepsilon \quad (10)$$

或者

$$\underset{\alpha}{\text{argmin}} \|y - \alpha X\|_2^2 + \lambda \|\alpha\|_1 \quad (11)$$

$y$  是需要重建的样本,  $X$  是嵌入向量的矩阵,  $\varepsilon$  和  $\lambda$  都是小的正常数。

尽管  $L_1$  范数在回归训练样本的选择中起着隐性作用, 但迭代解的计算成本很高, 用  $L_2$  范数代替正则化项。目标函数可以表示为

$$\underset{\alpha}{\text{argmin}} \|y - \alpha X\|_2^2 + \lambda \|\alpha\|_2^2 \quad (12)$$

稀疏表示为:

$$\underset{x_i}{\text{argmin}} \|k_i - x_i K\|_2^2 + \lambda \|x_i\|_2^2 \quad (13)$$

其中,  $\lambda$  是权重参数,  $k_i$  是测试文本中的第  $i$  个新单词对应的词向量,  $x_i \in R^m$  是重构向量,  $K \in R^{m \times n}$  由词库中的  $m$  个语义词向量组成,  $n$  是词向量维度, 最后将被稀疏表示过的新语义词添加到词库中, 用来提高模型的适应性。

### 1.3 文本特征提取

本文使用 BGRU 模型从训练的词向量组成的文本向量矩阵提取文本特征。

聊天记录中的“黑话”与原有的语义有较大的差别, 需要根据上下文才能理解“黑话”的歧义。本文在 BGRU 模型上进行改进来适应这一特点。BGRU 模型由两个反向的 GRU 模型组成, GRU 模型结构如图 3 所示。

图 3 中,  $Z_t$  为更新门,  $r_t$  为重置门,  $x_t$  表示  $t$  时刻的输入;  $h_{t-1}$  为隐藏层, 表示  $t-1$  时刻的输出;  $\sigma$  为 Sigmoid 函数;  $h_t$  为隐藏层, 表示  $t$  时刻的输出; 在  $x_t$  和  $h_{t-1}$  输入 GRU 之前分别添加固定值  $\beta$  和  $1-\beta$  作为权重来控制  $x_t$  和  $h_{t-1}$  的输入影响。GRU 模型中各个门计算如式(14)~式(17)所示。

$$z_t = \sigma(W_z \cdot [(1-\beta)h_{t-1}, \beta x_t]) \quad (14)$$

$$r_t = \sigma(W_r \cdot [(1-\beta)h_{t-1}, \beta x_t]) \quad (15)$$

$$h_t = (1-z_t) \cdot (1-\beta)h_{t-1} + z_t * \tilde{h}_t \quad (16)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \cdot (1-\beta)h_{t-1}, \beta x_t]) \quad (17)$$

其中,  $W_z$  表示更新门连接的权重矩阵,  $W_r$  表示重置门连接的权重矩阵, “ $\cdot$ ” 表示两个矩阵元素的相乘。

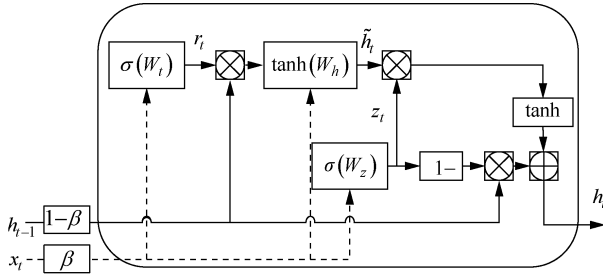


图3 GRU 模型结构  
Figure 3 Structure of GRU model

BGRU 模型结构如图 4 所示, 其中,  $\vec{h}_t$  表示  $t$  时刻的 GRU 的正向输出;  $\overleftarrow{h}_t$  表示  $t$  时刻的 GRU 的反向输出;  $h_t$  表示  $t$  时刻的 BGRU 的输出;  $x_t$  表示  $t$  时刻的输入。

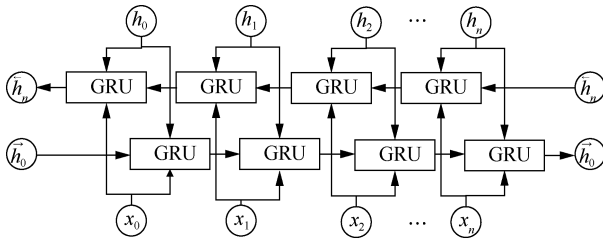


图4 BGRU 模型结构  
Figure 4 Structure of BGRU model

BGRU 模型中每一个时刻状态计算如式(18)、式(19)所示。输出则由这两个方向的 GRU 的状态共同决定, 如式(20)所示。

$$\vec{h}_t = \text{GRU}(x_t, \vec{h}_{t-1}) \quad (18)$$

$$\overleftarrow{h}_t = \text{GRU}(x_t, \overleftarrow{h}_{t-1}) \quad (19)$$

$$h_t = w_t \vec{h}_t + v_t \overleftarrow{h}_t + b_t \quad (20)$$

其中,  $w_t$  表示正向输出的权重矩阵,  $v_t$  表示反向输出的权重矩阵,  $b_t$  表示  $t$  时刻的偏置矩阵。

经过反复测试,  $\beta$  为 0.2 时, 分类效果最好, 这时, BGRU 模型提取的文本特征受到上下文

的影响较大, 这也符合聊天记录中的“黑话”特点。

#### 1.4 文本分类

本文通过 Keras 框架搭建 BGRU 模型来提取文本特征, 其中模型的组成部分有: 输入层使用 DSR 模型进行词的向量表示组成的文本矩阵; 大小为 64 的 BGRU 隐藏层; 从模型的两个方向输入的输入序列; 通过隐藏层提取文本的上文信息特征和文本的下文信息特征; 用式(21)结合隐藏层的两个方向。

$$h_{ijt} = \text{BGRU}(T_{ijt}) \quad (21)$$

其中,  $T_{ijt}$  表示在  $t$  时刻输入的第  $j$  个文本的  $i$  个词向量组成的文本矩阵;  $h_{ijt}$  表示在  $t$  时刻 BGRU 的输出。

输出层 softmax 的大小与文本分类的类别一致, 数据二分类后结果存在两个神经元, 分别代表正常与异常。

通过 softmax 函数进行文本分类, 分类函数如式(22)所示。

$$y_i = \text{softmax}(w_i h_{ijt} + b_i) \quad (22)$$

其中,  $w_i$  表示特征提取层到输出层的权重系数矩阵,  $b_i$  表示相应的偏置,  $h_{ijt}$  表示在  $t$  时刻特征提取层的输出向量。

## 2 实验分析

本文实验的目的如下。

1) 分析 DSR 模型中不同单词属性加权方法对分类效果的影响以及对新单词的稀疏表示方法的效果。

2) 分析犯罪领域的聊天记录存在的“黑话”现象特点。

3) 比较不同文本分类模型的性能。

### 2.1 实验环境

本文实验环境为 x86 平台, Intel CPU, 内存 16 GB, 硬盘 100 GB。操作系统 Windows10 家庭版, 使用 TensorFlow 的深度学习库 Keras 进行模型搭建与测试。

### 2.2 数据集

实验数据来自实验所用 Android 智能手机。手机中有与 1 000 个微信好友的聊天记录, 这 1 000 个会话内容包括正常聊天和与犯罪相关

的聊天。正常聊天和与犯罪相关的聊天主题如表 3 所示。本文将正常聊天内容标注为“正常文本”，与犯罪相关的文本标注为“异常文本”，这些聊天记录共包含 24 100 条聊天短文本。从聊天记录中随机抽取一些数据集，包括正常、异常两类，每个类别的数据集按 4:1 的比例分为训练集、测试集两部分。数据结果用 csv 格式进行存储，如表 4 所示。

表 3 聊天记录主题  
Table 3 The topic of chat record

| 文本类型 | 主题 |
|------|----|
| 正常文本 | 生活 |
| 正常文本 | 娱乐 |
| 正常文本 | 学习 |
| 正常文本 | 体育 |
| 正常文本 | 财经 |
| 异常文本 | 色情 |
| 异常文本 | 毒品 |
| 异常文本 | 赌博 |

表 4 数据集划分  
Table 4 Data set division table

| 语料类别 | 正常文本   | 异常文本   |
|------|--------|--------|
| 训练集  | 8 944  | 10 336 |
| 测试集  | 2 236  | 2 584  |
| 合计   | 11 180 | 12 920 |

另外，如表 5 所示，本文使用 4 个数据集来评估 DSR 模型的性能，其中 ChnSentiCorp\_htl\_all 来自谭松波的“酒店评论数据集”，waimai\_10k 是 Github 网站“某外卖平台收集的用户评价数据集”，online\_shopping\_10\_cats 是 Github 网站“各电商平台的用户评价数据集”，weibo\_senti\_100k 来自于新浪微博评论数据集。由于关于异常文本分类的数据集较少，因此本文挑选常见的用于文本情感倾向判别的中文数据集。数据集中 80% 为训练集，其余文本为测试集。

### 2.3 模型评估

本文采用精确率、召回率和  $F1$  值对分类结果进行评估。其中，FN (false negative)、FP (false positive) 和 TP (true positive) 分别表示正常文本预测为异常文本数量，异常文本预测为正常文本数量，正常文本预测为正常文本数量。计算如下所示。

$$\text{Precision} = \frac{TP}{TP+FP} \quad (23)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (24)$$

$$F1 = \frac{2PR}{P+R} \quad (25)$$

表 5 中文数据集  
Table 5 Chinese data set

| 数据集                     | 正面情感数量 | 负面情感数量 | 总数     |
|-------------------------|--------|--------|--------|
| ChnSentiCorp_htl_all    | 5 000  | 2 000  | 7 000  |
| waimai_10k              | 4 000  | 4 000  | 8 000  |
| online_shopping_10_cats | 3 000  | 3 000  | 6 000  |
| weibo_senti_100k        | 5 000  | 5 000  | 10 000 |

### 2.4 BGRU 模型超参数

实验采用参数固定法，确定 BGRU 模型的超参数，如优化器、损失函数、激活函数等。

本文使用的是 BGRU 模型训练超参数如表 6 所示。

表 6 BGRU 模型训练超参数  
Table 6 BGRU model training super-parameter

| 参数名称      | 参数值                |
|-----------|--------------------|
| 学习率       | $1 \times 10^{-3}$ |
| 优化器       | AdaDelta           |
| 损失函数      | MSE                |
| Dropout 率 | 0.5                |

经过反复实验测试，当采用表 6 的超参数，以准确率作为评估标准，训练轮数达到 27 时，实验效果最好。精确率随轮数变化如图 5 所示。

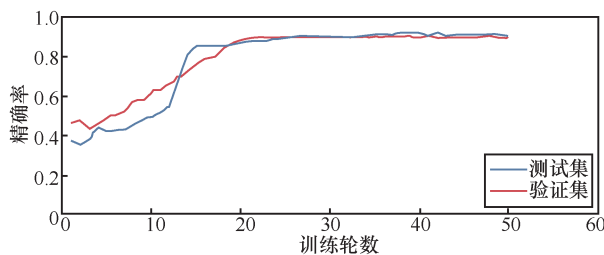


图 5 精确率随轮数变化  
Figure 5 The graph of accuracy with the number of rounds

### 2.5 DSR 模型性能评估

在 NLP 中有很多量化单词权重的方法，常见的有 TF-IDF 加权和单词频率加权。本文将单词的位置、频率、词性特征相结合来量化单词权

重。在实验中, 本文研究比较了这些量化单词权重方法的效果。图 6 中的结果表明, 加权组合比其他加权方法更有效。

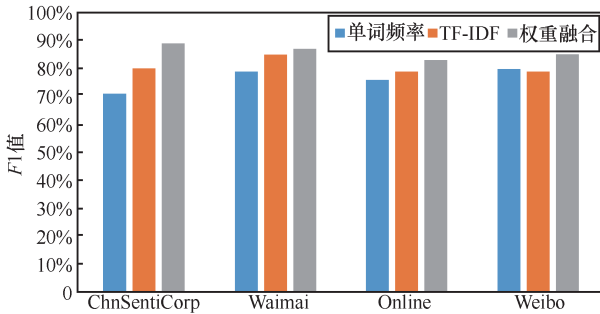


图 6 不同单词属性加权方法的影响

Figure 6 Influence of different word attribute weighting methods

为了验证稀疏表示方法的有效性, 本文在 4 个数据集上测试了去掉稀疏表示方法的情况下, 各方法的  $F1$  值。

如图 7 所示, 稀疏表示方法提高了大多数数据集上分类器的性能。ChnSentiCorp 数据集上的优化性能较差, 原因是训练集与测试集中的词重合比较大, 因此动态表示添加的新词较少。其他数据集上训练集和测试集词语重合较少, 但词义相近。这提高了稀疏表示的效果。

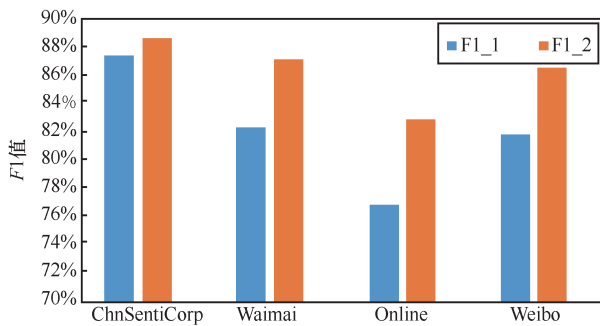


图 7 稀疏表示的影响

Figure 7 The influence of sparse representation

图 7 中  $F1\_1$  展示的是去掉稀疏表示的分类结果的  $F1$  值,  $F1\_2$  展示的是加上稀疏表示的分类结果的  $F1$  值。

## 2.6 “黑话”现象探讨

犯罪领域的聊天记录存在“黑话”现象。本文比较了被用作“黑话”的单词在数据集中正常文本和异常文本中的特征表示。由于单词的特征表示维度较高, 本文使用 sklearn 算法库中的 T-SNE 算法进行降维可视化, 结果如图 8 所示。

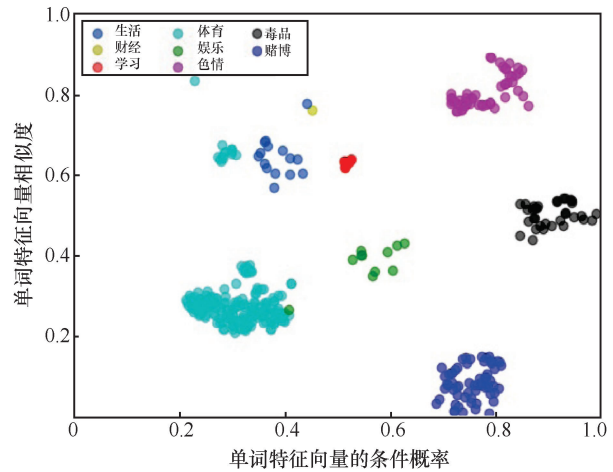


图 8 “黑话”足球在不同文本领域的特征表示

Figure 8 Feature representation of "slang" football in different text fields

如图 8 所示, “黑话”在异常文本领域中分布较广, 在正常文本中集中为特定领域。而且不同领域的“黑话”特征表示相近, 异常文本中不同领域的“黑话”特征表示相距较远。这说明“黑话”在犯罪领域中语义不固定, 且与原义有较大偏差。“黑话”的歧义问题也是自然语言处理中经典的“一词多义”问题, 所以利用上下文识别其真实意义是解决这一问题的方法。但在应用时应根据“黑话”的特点, 进行特征表示时减小词本身的含义影响, 放大上下文的影响。

## 2.7 模型对比

为了充分验证 DSR-BGRU 文本分类模型的有效性、增强对比性, 本文不仅与其他犯罪文本分类研究进行对比, 而且采取目前比较流行的面向社交媒体的文本分类模型进行比较分析。这些研究的详细信息如下。

1) DSR-BGRU: 本文提出的文本分类模型, 分别运用 DSR 模型和 BGRQ 模型对文本进行特征表示及提取。

2) 文献[2]: 文献[2]将 PageRank 算法和关系网络相结合, 提出一种 ARPR 算法。该算法采用 TF-IDF 方法提取群聊人员的群聊涉毒关键词, 并对涉毒关键词在涉毒嫌疑程度排序中的贡献进行度量; 然后以层次分析法引导聚合各维度信息计算得到的嫌疑人员权重为权重系数, 以好友关系为链接建立关系网络作为 PageRank 的入度与出度来计算相对应的 PageRank 权值。

3) 文献[14]: 文献[14]对单词的特征表示训

练方法进行了改进,根据分类任务目标对预训练的词嵌入基于对应领域的词库进行针对性的重构来增强分类效果。

4) 文献[15]: 文献[15]提出了 BLS (broad learning system) 模型用于文本分类,该模型在 LSTM 的基础上进行了改进,引入了注意力机制来关注重点单词,并且增加了遗忘门来控制上下文信息的传递。

5) 文献[16]: 文献[16]提出了一种被称为 DE-CNN 的神经网络,该网络可以将与上下文相关的概念合并到卷积神经网络中,以进行短文本分类。模型首先利用两层分别提取概念和上下文,使用关注层提取与上下文相关的概念;然后,将这些概念合并到文本特征表示中以进行短文本分类。

6) 文献[17]: 文献[17]使用 BERT 预训练语言模型对短文本进行句子层面的特征向量表示,并使用 softmax 对其进行分类。

由表 7 可以看出,本文提出的模型在准确率、召回率、 $F1$  这 3 个指标上都有较明显的优势。文献[2]采用的模型各项指标均最低,原因在于它采用 TF-IDF 方法提取群聊人员的群聊涉毒关键词,对于“黑话”的一词多义、更新速度快的特点难以适应,尽管是适用于分类涉毒领域的聊天文本,但在其他犯罪领域的聊天文本上表现较差。文献[14]与文献[2]在数据集上表现效果相近,是因为采用类似的方法强化在单一文本领域的表现能力,适应性较弱。文献[17]与文献[15]相比,性能较弱。是因为相比文献[17]引入的注意力机制,文献[15]根据分类目标对预训练的词嵌入能更增强模型的特征表示能力。文献[16]相比文献[15]和文献[17],将注意力机制结合卷积层神经网络,有效捕捉上下文语法语义以及更深层次的信息,从实验结果来看性能表现较好。

表 7 对比实验结果  
Table 7 Results of the compare experiment

| 模型       | 准确率    | 召回率    | $F1$ 值 |
|----------|--------|--------|--------|
| DSR-BGRU | 92.06% | 89.97% | 91.00% |
| 文献[2]    | 81.17% | 81.83% | 81.50% |
| 文献[14]   | 82.72% | 82.47% | 82.59% |
| 文献[15]   | 89.95% | 89.31% | 89.63% |
| 文献[16]   | 91.68% | 90.17% | 90.92% |
| 文献[17]   | 86.85% | 86.43% | 86.64% |

从对比实验的结果来看,能有效捕捉句子中的序列、上下文、语法语义以及更深层次的信息的模型在犯罪领域的聊天文本分类上有不错的表现能力。这是因为犯罪领域中“黑话”一词多义的特点要求模型从上下文语境中提取深层次信息的能力。本文提出的 DSR-BGRU 模型针对 BGRU 模型做的改进使得在犯罪领域的聊天文本分类上具有一定优势。

## 2.8 模型缺点

本节对数据集中分类错误的聊天文本进行了探讨。在分类错误的聊天文本中,表情符、特殊符号作为“黑话”来掩盖真实的犯罪意图。本文研究提出的模型在预处理环节丢弃了表情符和特殊符号,这可能会丢弃潜在的“黑话”,从而造成误判。

此外,本文针对 BGRU 模型做的改进是添加了固定参数  $\beta$  来控制输入单词与上下文对语义的影响。这种单一权重的使用降低了“黑话”单词对语义的影响,也降低了其他单词对语义的影响。

## 3 结束语

本文提出一个文本分类模型用于解决数字取证时遇到的聊天文本分类难题。本文提出 DSR 模型将语料中的词语用高维的词向量表示,接着将词向量输入改进后的 BGRU 模型中提取特征,最后连上一个 softmax 输出层,这样就建立了一个文本分类模型。该模型使用 DSR 对文本预训练来从语义层面对聊天文本进行特征表示,再使用改进后的 BGRU 对使用这些词向量组成的文本提取上下文特征,从而能够更好地准确理解评论文本的语义信息。实验结果表明,使用 DSR-BGRU 模型能够更加准确地完成识别和提取与犯罪事件有关聊天证据。

但 DSR-BGRU 模型存在一定不足,其无法处理聊天文本中的表情符和特殊符号以及使用固定参数控制输入单词与上下文对语义的影响。针对这些不足,可以对所有表情符号与特殊符号生成独特的特征向量后再引入文本分类中以及引入注意力机制<sup>[8]</sup>,根据单词词义或其他属性变化参数值来控制输入单词与上下文对语义的影响。

## 参考文献:

- [1] 中国互联网络中心. 中国互联网络发展状况统计报告[EB]. China Internet Network Center. Statistical report on Internet development in China[EB].
- [2] 刘昊, 徐鹏. 基于关系网络的 PageRank 算法在禁毒情报上的应用研究[J]. 中国人民公安大学学报(自然科学版), 2019, 25(1): 65-73.  
LIU H, XU P. Research on application of PageRank algorithm based on relational network in anti-drug intelligence [J]. Journal of People's Public Security University of China (Science and Technology), 2019, 25(1): 65-73.
- [3] 杜思佳, 于海宁, 张宏莉. 基于深度学习的文本分类研究进展[J]. 网络与信息安全学报, 2020, 6(4): 1-13.  
DU S J, YU H N, ZHANG H L. Survey of text classification methods based on deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(4): 1-13.
- [4] 于游, 付钰, 吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019, 5(5): 1-8.  
YU Y, FU Y, WU X P. Summary of text classification methods[J]. Chinese Journal of Network and Information Security, 2019, 5(5): 1-8.
- [5] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of Twenty-ninth AAAI Conference on Artificial intelligence. 2015.
- [6] TRAN K, BISAZZA A, MONZ C. Recurrent memory networks for language modeling[J]. 2016: arXiv: 1601.01272.
- [7] TANG D Y, QIN B, FENG X C, et al. Effective LSTMs for target-dependent sentiment classification[J]. 2015: arXiv: 1512.01100.
- [8] LIU G, GUO J B. Bidirectional LSTM with attention mechanism and convolutional layer for text classification[J]. Neurocomputing, 2019, 337: 325-338.
- [9] 赵宏, 王乐, 王伟杰. 基于 BiLSTM-CNN 串行混合模型的文本情感分析[J]. 计算机应用, 2020, 40(1): 16-22.  
ZHAO H, WANG L, WANG W J. Text sentiment analysis based on serial hybrid model of bi-directional long short-term memory and convolutional neural network[J]. Journal of Computer Applications, 2020, 40(1): 16-22.
- [10] 陈榕, 任崇广, 王智远, 等. 基于注意力机制的 CRNN 文本分类算法[J]. 计算机工程与设计, 2019, 40(11): 3151-3157.  
CHEN R, REN C G, WANG Z Y, et al. Attention based CRNN for text classification[J]. Computer Engineering and Design, 2019, 40(11): 3151-3157.
- [11] 张洋, 胡燕. 基于多通道深度学习网络的混合语言短文本情感分类方法[J]. 计算机应用研究, 2021, 38(1): 69-74.  
ZHANG Y, HU Y. Code-switching short-text sentiment classification method based on multi-channel deep learning network[J]. Application Research of Computers, 2021, 38(1): 69-74.
- [12] 丁建立, 苏现帅. 基于组合式深度学习网络的混合文本情感分类[J]. 计算机工程与设计, 2019, 40(11): 3254-3258, 3264.  
DING J L, SU X S. Mixed text classification method based on combined deep learning network[J]. Computer Engineering and Design, 2019, 40(11): 3254-3258, 3264.
- [13] 曹宇, 李天瑞, 贾真, 等. BGRU: 中文文本情感分析的新方法[J]. 计算机科学与探索, 2019, 13(6): 973-981.  
CAO Y, LI T R, JIA Z, et al. BGRU: new method of Chinese text sentiment analysis[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 973-981.
- [14] MOREO A, ESULI A, SEBASTIANI F. Word-class embeddings for multiclass text classification[J]. Data Mining and Knowledge Discovery, 2021, 35(3): 911-963.
- [15] DU J, VONG C M, CHEN C L P. Novel efficient RNN and LSTM-like architectures: recurrent and gated broad learning systems and their applications for text classification[J]. IEEE Transactions on Cybernetics, 2021, 51(3): 1586-1597.
- [16] XU J Y, CAI Y, WU X, et al. Incorporating context-relevant concepts into convolutional neural networks for short text classification[J]. Neurocomputing, 2020, 386: 42-53.
- [17] 段丹丹, 唐加山, 温勇, 等. 基于 BERT 模型的中文短文本分类算法[J]. 计算机工程, 2021, 47(1): 79-86.  
DUAN D D, TANG J S, WEN Y, et al. Chinese short text classification algorithm based on BERT model[J]. Computer Engineering, 2021, 47(1): 79-86.

## [作者简介]



张宇 (1996- ), 男, 江苏连云港人, 信息工程大学硕士生, 主要研究方向为智能手机取证。



李炳龙 (1974- ), 男, 河南卫辉人, 博士, 信息工程大学副教授, 主要研究方向为数字调查与取证、网络入侵溯源追踪与取证、云计算取证、智能手机取证。

李学娟 (1975- ), 女, 河南焦作人, 河南理工大学工程师, 主要研究方向为控制理论与控制工程。

张和禹 (1998- ), 男, 河南南阳人, 信息工程大学硕士生, 主要研究方向为内存取证。