改进 MacBERT 与双阶段相似度的警情识别与公安应用

李也桐[1] 刘致用[2]
LI Yetong LIU Zhiyong

摘 要 针对海量警情数据中相似案件识别效率低、中文文本语义复杂性高等问题，文章提出了一种基于改进 MacBERT 模型的相似警情识别方法。通过优化掩码语言模型任务，引入近义词替换、全词遮蔽及 N-gram 遮蔽策略，提升模型对中文复杂语法与警情领域术语的语义理解能力。进一步构建双阶段相似度计算框架，融合细粒度语义匹配和基于向量编码的快速检索方法，在保证高精度的同时提升计算效率。实验基于公开数据集 SimCLUE 与警情数据，对比 MacBERT 和文本嵌入模型，结果表明，改进 MacBERT 在损失值、准确率和 F_1 分数上均有良好表现，验证其对中文警情文本的强适应性；而文本嵌入模型凭借静态向量预训练特性，在运行时间与吞吐量上更具效率优势。实际公安场景验证表明，所提方法可高效支持案件串并分析、跨区域情报关联等任务，辅助缩短案件侦破周期，并优化警力资源配置。未来研究将聚焦多模态数据的语义融合与轻量化模型部署，进一步推动公安智能化转型。

关键词 相似警情识别；MacBERT；交互策略；文本嵌入；公安实战

doi：10.3969/j.issn.1672-9528.2025.09.030

0 引言

目前，公安机关掌握的警情数据种类繁多且颇具价值，并且其数量还在呈指数级增长[1]。在海量警情数据中，快速识别出相似的警情，既能帮助民警借鉴过往案例，提升整体办案效率，又能通过总结剖析相似案件经验，提炼出更为统一的执法标准和操作规范，提高执法公信力。同时，把握相似警情的特征，还有助于公安机关对新案件的复杂程度和所需警力做出更加准确地评估，并据此合理地利用警力资源。因此，对于公安机关而言，识别相似警情不仅是一项至关重要的任务，更是提升工作效能、增强执法公信力和优化资源配置的有力抓手。

相似警情识别的本质是对于文本相似度的识别，旨在确定两个或多个文本在内容或意义上的相似程度。被应用于各种场景，在教育系统中，帮助教师完成自动批改[2]；在审计过程中，解决审计师披露关键审计事项的问题[3]；在电气工程中，提出自主可控新一代变电站一键顺控测试方法[4]；在新闻媒体中，进行语义内容分析，探索媒介内容建设[5]；对文档的重复程度进行评估。中文文本的语义具有多样性和复杂性，也使得文本在相似度计算的任务中面临着巨大的挑战[6]。

1. 安徽公安学院 安徽合肥 238000
2. 河北石油职业技术大学 河北承德 067000

警情文本作为案件记录的核心载体，通常包含报警人描述、案情细节及民警处置信息。然而，由于不同地区警务规范的差异性、民警表述习惯的多样性，以及报警人或犯罪嫌疑人语言表达能力与心理状态的波动性，导致文本中关键语义特征提取面临挑战。为此，本文提出一种基于改进 MacBERT 模型的混合式相似度计算方法。

1 基于 Transformers 的相似警情识别模型

1.1 模型结构

MacBERT 是一种创新的深度学习模型，在 BERT 的基础上进行了针对性的优化，旨在解决 BERT 在掩码语言模型任务上的不足[7]，并且使用了约含 172 000 个汉字中文词汇表和 12 层 Transformer 编码器结构，有利于学习复杂语法，能更好地适应中文数据。MacBERT 保留了 BERT 的基本框架，但对 MLM 任务进行了修正，引入近义词替换和 Ngram 掩码策略，从而减少预训练和微调阶段之间的任务差异，提升模型的性能。

1.2 基于 MacBERT 的警情相似度计算方法

1.2.1 两条文本的相似度识别

把句子对输入到预训练的模型中，对其进行全连接处理，进而判断两个文本之间的相似度。输入处理与特征编码，将待比对的两个文本 A、B 输入改进的 MacBERT 模型中。提取上下文语义特征，利用 MacBERT 的 12 层 Transformer 编码器对词向量序列进行多尺度上下文建模[8]。捕捉文本间的

语义关联性，引入交叉注意力机制构建交互特征。并将交互特征输入到全连接层中，对交互特征进行非线性变换，提取更高层次的特征。最后通过全连接层的输出，计算两个文本之间的相似度得分。

1.2.2 多条文本的相似度识别

处理多条文本的相似度识别时，将上述基于交互策略的两条文本相似度计算方法进行扩展。具体步骤如下：

（1）输入处理与特征提取：对每一条警情描述文本进行输入处理和特征提取，得到每个文本的特征向量。

（2）构建文本特征矩阵：将所有文本的特征向量组合成一个特征矩阵，其中每一行代表一个文本的特征向量。

（3）计算相似度矩阵：利用特征矩阵中的每一对特征向量，计算它们之间的相似度得分。得到一个相似度矩阵，其中每个元素代表两个文本之间的相似度得分。

（4）相似度排序与筛选：对于每个文本，根据相似度矩阵中的得分，对其余文本进行相似度排序。根据实际需求，设定一个相似度阈值，筛选出与当前文本相似度高于阈值的文本。

（5）结果输出：输出每个文本与其相似文本列表，以及相应的相似度得分。

1.3 基于文本嵌入的警情相似度计算方法

将已存在的候选文本通过向量编码模型转变为向量表示，映射到高维空间中，存到向量数据库中[9]。有新文本输入时，将其向量表示后，去向量库中进行向量匹配，衡量新文本与已存在的候选文本之间的相似程度。利用预训练的向量编码模型，将警情文本中的句子转化为向量表示。每个向量代表文本中的一个特征，向量的维度通常较高，以捕捉文本的细微差别。计算新文本向量与候选文本向量之间的相似度。并用损失函数评估两个向量之间的相似性，根据相似度计算结果，输出与新文本相似的候选文本列表或相关警情信息。

2 实验

2.1 数据集与实验环境

2.1.1 数据集

本文共采用两个数据集：

（1）SimCLUE 数据集：该数据集整合了中文领域绝大多数可用的开源的语义相似度和自然语言推理的数据集，共 2 678 728 条，并重新做了数据拆分和整理。本文采用 JSON 格式的数据集，训练集和测试集按照 8:2 分配。

（2）某省某市局警情数据：共 30 000 条。

2.1.2 实验环境

本次实验是在 Miniconda 的虚拟环境中完成的，使用 PyTorch 中有专门的模块 torch.cuda 来设置和运行 CUDA 相关操作。本地安装环境为 Windows10、Python3.9.20、CUDA 11.6 和 Transformers 4.46.3，编辑器使用 vscode 并通过其 Jupyter 扩展来运行 Notebook。

2.2 评价指标

本次实验采用的评价指标：

（1）loss：模型损失值，衡量模型预测与实际结果的差异；

（2）accuracy：模型准确率，正确预测的数量除以总样本数；

（3）F_1-score：精确率和召回率的调和平均数；

（4）Runtime：模型在数据集上运行所需时间；

（5）Samples Per Second：模型每秒可以评估的样本数；

（6）Steps Per Second：在批处理的情况下，每秒可以完成的批处理步骤；

（7）Epoch：训练周期，一个训练周期表示整个训练数据集被模型学习了一次。

2.3 实验设置与结果分析

2.3.1 实验设置

为验证 MacBERT 模型在文本相似度分析任务中的有效性，本文将 MacBERT 模型与文本嵌入模型进行比较，实验流程如图 1 所示。

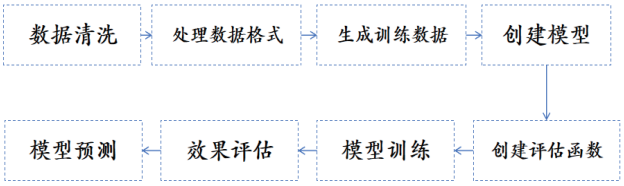

图 1 实验流程图

（1）数据清洗：主要收集并处理原始的警情文本数据。

（2）处理数据格式：清洗后的数据需要进一步处理成适合深度学习模型训练的格式。

（3）生成训练数据：为每对文本分配一个相似度标签。在利用警情数据做实验时，将分数大于 70% 的数据评定为相似文本。

（4）创建模型：本次实验主要对比 MacBERT 和文本嵌入模型的相似度识别。

（5）创建评估函数：评估函数用于衡量模型在测试集上的性能。

（6）模型训练：使用训练集数据对模型进行训练。

（7）效果评估：在验证集上使用评估函数对训练好的模型进行性能评估。

（8）模型预测：使用训练好的模型对新的警情文本进行相似度预测。

2.3.2 结果分析

图 2 为 MacBERT 模型的 loss 值，图 3 为文本嵌入模型的 loss 值。

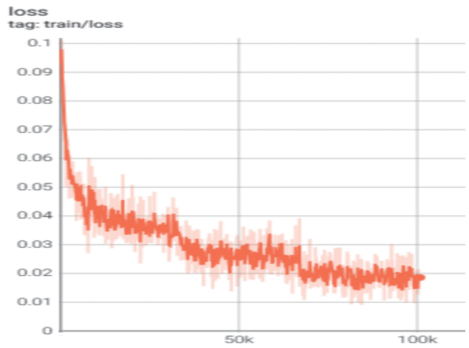

图 2 MacBERT 模型损失值变化曲线

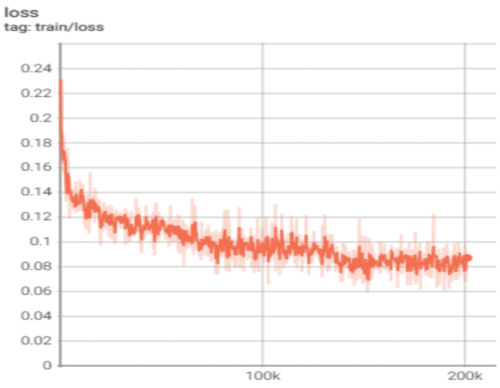

图 3 文本嵌入模型损失值变化曲线

表 1 展示出 MacBERT 的 loss 值优于文本嵌入模型，准确率也比较高，主要优势来源于通过改进 MLM 修正策略，显著提升了模型对中文语义的捕捉能力。

表 1 对比实验

模型	loss	accuracy/%	F_1-score/%	Runtime /min	samples per second	steps per second	Epoch
MacBERT	0.031 8	96.13	95.15	997.225	537.237	10.6	3.0
文本嵌入	0.091 5	89.87	87.12	789.822	678.312	16.79	3.0

对比实验中文本嵌入模型的运行时间更短，究其原因，是预训练与微调策略，文本嵌入模型通常基于预训练好的静态向量，无需针对特定任务进行端到端微调，节省了训练与推理时间。并且向量生成后可直接存储，支持离线批量处理，适合大规模相似度检索。而 MacBERT 模型需在特定任务上微调，在推理时还需逐样本动态编码，增加了实时计算负担。MacBERT 的交互策略需对每对文本进行联合编码，导致计算量成倍增长。通过本实验得出，MacBERT 在警情文本的深度语义理解上表现卓越，适合高精度需求场景；文本嵌入模型则以效率见长，适用于实时性与资源敏感任务。实际应用中需根据具体需求权衡精度与效率，或通过技术融合实现优势互补。

3 实战应用

该技术可扩展至案件串并分析、舆情监测、情报关联研判[10]等公安核心场景，如表 2 所示，助力提升案件侦破效率、舆情预警能力与执法规范化水平。

表 2 公安领域应用场景

应用场景	技术适配	效益
案件串并分析	使用 MacBERT 提取案件文本深层语义特征，结合交互策略与相似度阈值筛选高关联案件。	缩短案件侦破周期，提升对惯犯或团伙犯罪的打击效率。
情报关联与研判	构建情报特征库，利用向量匹配技术实时关联新情报与历史数据。	辅助构建犯罪网络图谱，提升反恐、缉毒等专项工作的情报支持能力。
舆情监测与危机预警	MacBERT 聚类相似舆情文本，结合情感分析触发高危事件预警。	实现早发现、早干预，降低社会不稳定风险。
嫌疑人行为模式分析	交互策略分析通信记录相似度，向量匹配关联嫌疑人行为轨迹。	辅助预判嫌疑人行动意图，提升抓捕成功率。

示例 1：案件串并分析

输入警情：“某商场发生金店抢劫案，嫌疑人戴口罩、持匕首，驾驶无牌摩托车逃离。”

系统输出：匹配到 3 起相似案件（相似度 >90%），均发生在相邻区县，嫌疑人特征高度一致。自动生成串并案建议，并在地图标记逃逸路径，提示重点布控区域。

示例 2：舆情预警

输入文本：社交媒体帖子“某小区连续发生多起入室盗窃，物业不作为！”

系统输出：关联近期 5 起相似报警记录，触发舆情预警。推送至辖区民警核实，同步生成舆情报告供指挥中心研判。

4 小结与展望

本研究针对中文警情文本的复杂性，对 MacBERT 的 MLM 任务进行改进，引入近义词替换与全词、N-gram 遮蔽策略，提升了模型对中文语法和领域术语的语义捕捉能力。并提出双阶段相似度计算框架，通过交互策略实现高精度语义匹配，结合向量匹配方法支持大规模快速检索。

当前研究依赖文本数据，后续可探索多模态相似度分析，融合图像、语音等非结构化数据，增强复杂案件的综合分析能力。针对实际部署需求，需优化模型推理效率，并加强数据隐私保护机制。

面向无监督跨域开集场景的动态熵阈值辐射源个体识别方法

王 闯[1] 俞 璐[1]
WANG Chuang YU Lu

摘 要 辐射源个体识别（specific emitter identification, SEI）技术在军事和民用领域应用广泛，在多数情况下，训练和测试数据并不满足独立同分布假设，且往往测试数据包含比源域更多类别，为了解决这一问题，文章提出了一种面向无监督跨域开集场景的动态熵阈值辐射源个体识别方法。该方法引入最大均值差异（maximum mean discrepancy, MMD）以减小训练数据和测试数据之间的差异，解决训练和测试数据分布不同导致的模型性能降低的问题。将分类器输出各类别熵的平均值作为未知类鉴别的阈值，并通过移动平均方式更新各类别平均熵，以动态获得熵阈值。实验结果表明，所提方法在 Oracle 和 Wisig 射频指纹数据集上均具有较高的识别准确率，验证了其有效性。

关键词 辐射源个体识别；开集域适应；熵阈值；最大均值差异

doi：10.3969/j.issn.1672-9528.2025.09.031

0 引言

辐射源个体识别（specific emitter identification, SEI）技术，也称特定辐射源个体识别技术，是一种通过对辐射源发射的电磁信号进行分析，提取其特有的特征，从而实现对辐射源个体身份识别的技术。SEI 技术在军事和民用领域均具有极大应用价值。不同辐射源个体，由于其制造工艺差异、元件特性以及使用过程中的衰损等因素，会形成独特的信号特征，这些特征被称为“射频指纹”。通过对接收信号数据的分析，提取这些指纹特征，并结合先验知识，即可实现对辐射源个体的识别。

近年来，随着人工智能技术的迅猛发展，深度学习在图像识别、图像分割、自然语言处理等领域得到广泛应用，并

1. 陆军工程大学 江苏南京 210000

[基金项目] 国家自然科学基金资助项目“复杂对抗条件下的电磁信号高可靠智能识别方法研究”（62471486）

参考文献：

[1] 张静 , 高子信 , 丁伟杰 . 基于 BERT-DPCNN 的警情文本分类研究 [J]. 数据分析与知识发现 , 2025,9(2):48-58.

[2] 赵聚雪 . 基于短文本相似度分析的测试用例自动批改研究 [J]. 电脑编程技巧与维护 , 2023(12): 127-129.

[3] 王西子 , 李英 , 吴联生 . 客户重要性与关键审计事项披露 : 基于关键审计事项披露数量与文本相似度的分析 [J]. 会计研究 ,2023(12):159-173.

[4] 杨宏伟 , 张红梅 , 张骥 , 等 . 基于 TF-IDF 加权文本语义相似度算法的变电站一键顺控测试方法研究 [J]. 电力科学与技术学报 , 2023,38(5):269-278.

[5] 高彦婷 . 新型主流媒体中新闻文本和评论的语义相似度分析 : 基于上观新闻语料库的 LDA 主题建模 [J]. 新闻传播 , 2023(11):18-20.

[6] 李莹 , 伍胜 , 徐聪 , 等 . 语义文本相似度计算方法研究综述 [J]. 软件导刊 , 2024, 23 (11): 1-11.

[7] 蒋晨 . 基于 BERT 模型的电气设备相似文本检索与故障聚类识别研究 [D]. 北京 : 华北电力大学 ,2022.

[8] 张琳琳 , 杨雅婷 , 陈沾衡 , 等 . 基于深度学习的相似语言短文本的语种识别方法 [J]. 计算机应用与软件 , 2020, 37(2): 124-129.

[9] 贺益侗 . 基于 doc2vec 和 TF-IDF 的相似文本识别 [J]. 电子制作 , 2018(18): 37-39.

[10] 郑建波 . 警情信息识别学习模型应用研究 [J]. 消防界 (电子版), 2023, 9(9): 51-53.

【作者简介】

李也桐（1998—），女，山东德州人，硕士研究生，助教，研究方向：自然语言处理、公安大数据等。

刘致用（2005—），男，河北沧州人，本科在读，研究方向：大数据、自然语言处理。

（收稿日期：2025-03-20 修回日期：2025-09-03）