Software Engineering and Applications 软件工程与应用, 2025, 14(4), 960-965
Published Online August 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.144085

Hans汉斯

基于BERT模型的保密管理应用研究综述

张　羽

中国航天科工集团第十研究院，贵州　贵阳

收稿日期：2025年7月7日；录用日期：2025年8月19日；发布日期：2025年8月28日

摘　要

在数字化转型不断深化的当下，保密管理面临着档案分类需求、敏感信息提取以及文本隐藏等多重核心挑战。作为自然语言处理领域的革命性突破，基于Transformer架构的BERT预训练语言模型能够深层次训练出包含上下文特征信息的语境化词向量，为上述挑战提供了新思路。本文聚焦BERT模型在保密管理中敏感信息处理领域的应用，探讨了BERT模型如何通过智能分类优化档案管理，结合实体识别和关系抽取等技术实现敏感信息的精准提取以及在文档中嵌入不可见的数字水印。研究表明，BERT预训练语言模型在文本敏感信息处理领域具有广阔的应用空间，其相关技术为保密管理提供了高效、精准的技术支持。

关键词

BERT模型，保密管理，档案分类，敏感信息提取，文本隐藏

A Review of BERT-Based Models for Confidentiality Management Applications

Yu Zhang

The 10th Research Academy of CASIC, Guiyang Guizhou

Received: Jul. 7th, 2025; accepted: Aug. 19th, 2025; published: Aug. 28th, 2025

Abstract

In the context of deepening digital transformation, confidentiality management faces multiple core challenges such as file classification requirements, sensitive information extraction and text hiding. As a revolutionary breakthrough in the field of natural language processing, BERT pre-trained language model based on the Transformer architecture is able to deeply train contextualized word vectors containing contextual feature information, which provides a new idea for the above challenges. This paper focuses on the use of BERT model in confidentiality management. This paper focuses on the application of BERT model in the field of sensitive information processing in confi-

文章引用: 张羽. 基于 BERT 模型的保密管理应用研究综述[J]. 软件工程与应用, 2025, 14(4): 960-965.
DOI: 10.12677/sea.2025.144085

dentiality management, and explores how BERT model can optimize the file management through intelligent classification, combine the techniques of entity recognition and relationship extraction to realize the accurate extraction of sensitive information as well as the embedding of invisible digital watermarks in documents. The study shows that the BERT pre-trained language model has a broad application space in the field of text sensitive information processing, and its related technology provides efficient and accurate technical support for confidential management.

Keywords

BERT Model, Confidentiality Management, File Classification, Sensitive Information Extraction, Text Hiding

1. 引言

在数字化转型加速推进的当下，重大工程项目、科研机构及企业面临的敏感信息保护挑战愈发严峻。随着电子文档与网络通信的广泛应用，海量数据高频交互，信息泄露风险显著加剧。传统保密管理体系主要是物理边界的防御模式为主，在虚拟空间安全威胁面前逐渐暴露出局限性，难以有效应对数字化场景下的动态变化。构建适配复杂数字生态的新型保密机制，已成为维护国家安全、保障企业竞争优势及社会稳定的关键命题。

随着人工智能的快速发展，自然语言处理技术为破解传统保密管理困境提供了新思路。档案分类[1]作为保密信息管理的基础架构，通过结构化组织实现信息的快速检索与权限控制。传统档案分类主要是以部门或职能划分，这种模式难以适应跨领域项目的复杂需求。例如在大型工程中，设计图纸、技术参数与供应商合同往往涉及多个部门协同，缺乏语义关联的分类体系容易形成信息孤岛，导致访问控制机制难以精准识别数据权限。在敏感信息[2]防护层面，传统的关键词识别方法已无法有效捕捉文本中的隐含语义关联，存在敏感词提取精度不足的问题。基于预训练语言模型的实体识别与关系抽取技术，通过双向上下文建模能力，能够更深入理解复杂语言信息，提升敏感信息识别准确率，为后续定密管理等下游任务提供可靠支撑。此外，文本隐藏术(如隐写术) [3]为信息溯源与防拍屏提供了创新的解决方案。通过在文档中嵌入不可见的数字水印，可实现泄露源头追溯至具体设备和操作人员，为构建全链路保密管理体系提供技术保障。

相较于传统 Word2Vec [4]等静态词向量模型，BERT [5]预训练语言模型结构基于 Transformer，能够捕捉文本深层特征与长距离依赖关系，在问答系统、文本分类等任务中展现出显著优势，也为保密技术革新注入核心动力。基于上述相关背景，本文首先对 BERT 预训练语言模型的相关概念进行阐述，围绕文档分类、敏感信息提取和文本隐藏三个方向对现有的基于 BERT 预训练语言模型的相关技术研究进行了归纳总结；最后，对该领域的未来发展趋势进行了展望。

2. BERT 模型概述

为捕捉文本上下文对词汇语义的影响，2018 年 Devlin 等人[5]提出了 BERT 预训练语言模型。该模型基于 Transformer 架构，通过大规模无监督预训练学习丰富的语言表示，结合双向上下文信息，BERT 能够同时考虑文本中词汇前后的语义线索，从而实现对文本语义的深度理解。在预训练阶段，BERT 采用

掩蔽语言模型(Masked Language Model, MLM)任务，通过随机掩蔽输入序列中 15%的词汇，从而使模型能够学习深层次的双向上下文特征；同时引入下一个句子预测(Next Sentence Prediction, NSP)任务，通过判断句子间的连贯性，帮助模型捕捉句子之间的逻辑关系。大量实验表明，BERT 在自然语言处理任务中展现出显著的性能优势。BERT 模型结构图如图 1 所示：

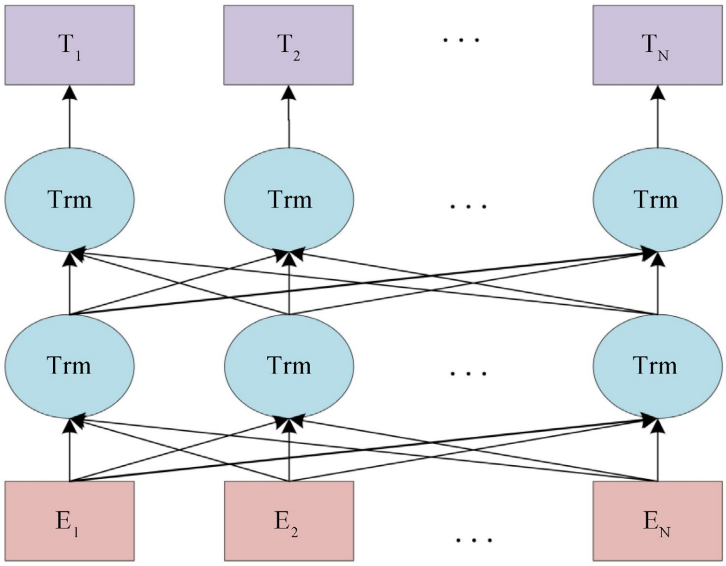

Figure 1. BERT model structure diagram
图 1. BERT 模型结构图

BERT 模型采用微调(Fine-tuning)方法实现模型优化，其核心在于通过预训练阶段构建深层语义表征能力，并针对具体的下游任务进行适配。这些下游任务涵盖命名实体识别、情感分析、文本分类等多个领域。在实际应用中，通过引入特定任务的数据集对 BERT 进行微调，动态调整模型的权重参数，使下游任务模型能更深入地挖掘上下文语义信息，进而生成更精准的词向量表征，最终实现任务处理效能的显著提升。

3. BERT 模型在保密管理中的相关技术研究

3.1. 档案分类

档案分类作为保密管理的关键存档环节，通过构建结构化数据组织框架，为敏感信息的分级保护与权限划分管控奠定了基础。随着数据交互频率与复杂程度的提升，传统的档案分类模式逐渐暴露出语义关联不足、跨领域适应性差等问题，但 BERT 模型通过其强大的预训练语言表示能力，能够有效地捕捉文本中的语义信息，在档案分类任务中展现出了显著的优势，从而提高分类的准确性。刘等人[6]引入 MacBERT 模型获取文本的动态特征表示，采用多尺度融合网络捕获档案文本局部特征和全局语义特征，有效解决了多义敏感词表征问题，以快速正确识别档案的类别；在特征提取层面，刘等人[7]提出了一种融合 ALBERT 与多通道特征网络的分类模型，通过并行捕获词级、短语级和篇章级的语义特征，结合注意力机制实现特征权重动态分配；肖等人[8]采用 BiLSTM-CRF 模型对文本关键词进行提取，构建关键词库，通过引入 Sentence-BERT 模型计算文本间的相似度，并与档案文件进行比对，以构建档案文件齐全性检验系统。这些改进方案推动了 BERT 模型在档案分类领域的深化应用，从静态词向量到动态语境建模，从单尺度特征提取到多通道特征融合的方法，不仅显著提升了分类准确率，更为档案分类管理提供

了完整的技术解决方案。

3.2. 敏感信息提取

敏感信息识别与提取是定密工作等保密管理中的核心环节，其关键在于精准提取敏感信息词，以此判定信息密级并进行自动化标注。传统的定密工作主要依赖人工标注，然而，面对海量异构文本时，人工标注存在效率低或者标注不一致等问题。为解决这一难题，基于 BERT 的预训练语言模型为敏感信息提取提供了有效技术支持。曾等人[9]提出一种 BERT-BGRU-CRF 模型，构建了层级化特征提取框架。该模型首先利用 BERT 预训练层获取动态词向量表征，通过双向门控循环单元(BiGRU)捕捉长距离语义依赖关系，最终通过条件随机场(CRF)实现密级标签的序列化标注。实验表明，该模型在涉密敏感信息识别任务中 F1 较传统方法提升 4.03%；在多标签分类领域，陆[10]提出一种基于 BERT-TextCNN 的多模态分类方案。针对开源威胁情报文本特点，设置正则判断规则，对标题和正文分别设置 Bert-TextCNN 多标签分类模型，将两部分标签整理去重得到文本的类别，该模型在 F1 指标上较对比模型提升了 2.02%，有效帮助安全研究人员高效率地对海量开源威胁情报文本信息进行筛选和浏览；面对网络言论的语义复杂性和隐晦性，闫等人[11]对文本卷积神经网络 TextCNN 进行了改进，结合 ALBERT 动态字级编码模型、多头自注意力机制与门控机制的优势，提出了一种融合字词特征的双通道分类模型 ALBERT-CCMHSAG，将文本与词级语义信息、局部关键特征与上下文语义进行了提取与融合，以此提升敏感言论的分类效果；为解决长文本处理中的信息衰减问题，李等人[12]提出一种融合敏感关键词特征的 Mer-Hi-Bert 模型，该模型通过分块模块将文档级别的互联网新闻分成若干个片段，对每个片段使用层次化 BERT 模型提取语义特征，进而使用注意力网络对关键片段进行加权融合，以此提高对敏感信息识别的精确率；针对网络安全实体稀疏性的特点，林等人[13]构建了一种基于语义增强的网络安全实体识别的模型，该模型采用双向长短期记忆模型(BiLSTM)获取上下文语义的隐藏表征，然后利用多头注意力机制对关键实体进行权重标定，最终通过 CRF 层生成最优标注序列；高等人[14]提出一种改进的 BERT-CRF 模型，采用双向 Transformer 编码器增强模型对上下文语义敏感特征的捕获能力，并引入鲸鱼优化算法(WOA)对模型进行超参数自适应调优，从而提高模型对敏感信息的识别精度和效率。

3.3. 文本隐藏

敏感信息隐藏是保密通信的关键技术，其关键在于在开放载体中嵌入秘密信息的同时，确保语义不可感知性。传统隐写术主要分为三类技术路径[3]：第一类是基于文本修改的方法，通过调整字符间距、替换同义词等方式嵌入信息，但存在隐写容量有限、易被统计检测算法识别等固有缺陷；第二类是基于文本选择的方法，依赖预先构建的载体库进行信息映射，但在长文本应用中易导致上下文断裂和语义漂移；第三类是基于文本生成的方法，虽然能生成逻辑连贯的文本，但早期基于 LSTM 的模型存在语义控制能力弱、生成文本可疑度高等技术瓶颈。

随着深度学习技术在自然语言处理领域的突破性进展，基于语言模型(LM)的隐写文本生成方法逐渐成为研究热点。此类技术通过预训练语言模型生成承载秘密信息的文本载体，以此提升了信息嵌入容量。Zheng 等人[15]针对现有方法的安全性缺陷，提出一种基于 BERT 与一致性编码的新型自回归语言模型(AR-LM)隐写算法，引入掩码语言模型机制，通过一致性编码技术突破传统块编码方法的局限性，实现对任意规模候选标记集的动态编码，并利用概率分布特性完成信息隐藏，增强了上下文语义的关联性；Ueoka 等人[16]提出一种利用掩码语言模型的解决方案，将 BERT 模型引入文本隐写框架，通过替换掩码位置实现隐写文本生成。Ding 等人[17]构建了融合 BERT 掩码语言模型与图注意力网络(GAT)的联合隐写模型，该模型通过多头自注意力机制与图注意力机制的协同作用，在文本生成过程中同步提取时序特征与空间

特征，显著提升了隐写前后文本特征的连贯性，进而增强了语言隐写模型的统计不可见性。通过对比实验和消融实验结果表明，联合语言隐写模型在隐蔽性和嵌入容量方面均优于其他最先进语言隐写模型。

3.4. 分类与比较：BERT 相关模型及其特点

基于上述文献中 BERT 相关模型在结构设计上的关键差异及其典型应用场景，可将其主要分为六类：(1) 基础 BERT 变体，(2) BERT 与序列模型结合，(3) BERT 与 CNN 结合，(4) BERT 与层次化/分块模型结合，(5) BERT 与优化算法结合，(6) BERT 与生成/隐写模型结合。表 1 从多个维度对这些模型的特性进行了详细比较。

Table 1. Classification and multidimensional comparison of BERT related models
表 1. BERT 相关模型分类及多维度比较表

模型类别	BERT 变体	BERT 与序列模型结合	BERT 与 CNN 结合	BERT 与层次化/分块模型结合	BERT 与优化算法结合	BERT 与生成/隐写模型结合
典型模型示例	MacBERT、ALBERT、Sentence-BERT	BERT-BGRU-CRF	BERT-TextCNN、ALBERT-CCMH-SAG	Mer-Hi-Bert	WOA-BERT-CRF	BERT-AR-LM、BERT-GAT
性能	BERT 变体提升多义词表征能力，F1 值优于传统 CNN	F1 值较传统方法提升，擅长密级标签序列化标注	多标签分类 F1 提升，敏感言论分类效果较好	提升长文本敏感信息识别精确率，解决信息衰减问题	提升敏感信息识别精度和效率，超参数自适应调优	增强上下文语义关联性，隐写文本统计不可见性提升，优于传统方法
优缺点	优点：轻量化、训练效率高，保留 BERT 核心语义理解能力；缺点：复杂语义任务中略逊于深层 BERT	优点：捕捉长距离语义依赖，序列标注精准；缺点：训练耗时，对短文本优势不明显	优点：融合局部关键特征与上下文语义，适合多标签任务；缺点：长文本局部特征易丢失	优点：擅长长文本处理，关键片段加权融合精准；缺点：分块策略影响性能，结构复杂	优点：模型参数优化更高效，识别精度提升；缺点：WOA 迭代可能陷入局部最优	优点：隐写安全性高，语义不可感知性强；缺点：生成文本逻辑一致性需优化
使用场景	档案分类、文本相似度计算、敏感言论初步筛选	涉密敏感信息识别、实体序列化标注显	开源威胁情报多标签分类、网络敏感言论识别	互联网新闻等长文档敏感信息识别	敏感信息高效识别	敏感信息隐藏、隐写文本生成

4. 结论

随着自然语言处理技术的快速发展，基于深度学习的敏感信息处理技术已成为保密管理领域的主流方法，其中 BERT 模型凭借其强大的语义表征能力展现出重要应用价值。该模型不仅有效提升了档案分类精度，优化了敏感信息提取准确率，还通过增强文本隐写技术的安全性，为保密管理提供了高效精准的技术支撑。本文系统综述了 BERT 模型在保密管理文本信息处理中的关键技术进展，重点探讨了档案分类中的动态语义建模技术、敏感信息语义特征提取方法，以及文本隐写术中掩码语言模型与注意力机制的协同优化策略。

然而，在复杂的现实应用场景中，BERT 模型仍面临多重挑战：其一，跨机构协作训练中的梯度更新过程存在潜在的信息泄露风险；其二，模型对专业领域术语的上下文关联特征捕获能力不足，可能导致定密不准确等问题；其三，模型决策过程的可解释性难以满足保密管理的规范要求。针对上述挑战，未来的研究可围绕以下方向进行系统性研究：

1) 强化隐私保护机制：重点开发基于差分隐私的 BERT 微调框架，设计自适应噪声注入机制，在保

障模型性能的同时实现训练数据的隐私保护；或构建联邦学习与 BERT 的融合架构，设计合适的梯度聚合协议，支持跨机构敏感文档的协同建模。

2) 增强专业领域理解：结合保密领域知识图谱增强模块，引入卷积神经网络(CNN)、图卷积神经网络(GCN)等技术，强化专业术语的上下文特征提取能力；通过实体链接与关系推理深化语义理解，并探索根据文档密级动态调整模型深度与注意力头数量的机制。

3) 提升模型可解释性：建立注意力权重校准机制，利用规则约束使关键语义区域的注意力分布更符合保密规范。

当前，保密管理领域缺乏公开可用的基准数据集，这在很大程度上制约了相关技术方案的客观评估与横向对比。因此，亟需构建涵盖党政机关、军工企业、科研院所等典型场景的标注文档的专用保密管理基准数据集。鉴于 BERT 模型在该领域展现的应用潜力，建立适配保密场景特性的 BERT 模型评估标准，重点考虑模型在安全性、准确性及可解释性等核心维度的性能表现。在此基础上，可推动 BERT 模型在国家级涉密单位实施试点应用，实现精准定密，最终构建可推广的智能化保密管理解决方案。

参考文献

[1] 孙桂玲, 郭振永. 军工单位保密工作档案管理的实践与思考[J]. 保密工作, 2022(11): 49-50.

[2] 吴海燕, 刘惠聪, 杨淑慧, 等. 企业涉密及敏感信息管理与业务融合研究——以电网企业档案工作为例[J]. 浙江档案, 2025(6): 51-54.

[3] 施昊, 郭文普, 陈昊. 基于深度学习的生成式文本隐写研究综述[J]. 火箭军工程大学学报, 2025, 39(2): 124-136.

[4] Mikolov, T., Sutskever, I., Chen, K., *et al*. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **2**, 3111-3119.

[5] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810. 04805.

[6] 刘舸舸. 结合 MacBERT 和多尺度融合网络的档案数据分类研究[J]. 电子设计工程, 2022, 30(19): 65-68+73.

[7] 刘景霞. 融合 ALBERT 与多通道特征网络的档案数据分类模型[J]. 电子设计工程, 2023, 31(15): 6-10.

[8] 肖雪丽, 廖常辉, 李惠仪. 一种基于深度学习的档案文件齐全性检验方法[J]. 信息记录材料, 2024, 25(3): 198-200+204.

[9] 曾庆瑞. 基于深度学习的涉密敏感信息识别技术研究[J]. 现代信息科技, 2024, 8(11): 171-175.

[10] 陆佳丽. 基于 Bert-TextCNN 的开源威胁情报文本的多标签分类方法[J]. 信息安全研究, 2024, 10(8): 760-768.

[11] 闫尚义, 王靖亚, 朱少武, 等. 融合字词特征的互联网敏感言论识别研究[J]. 计算机工程与应用, 2023, 59(13): 129-138.

[12] 李姝, 张祥祥, 于碧辉, 等. 互联网新闻敏感信息识别方法的研究[J]. 小型微型计算机系统, 2021, 42(4): 685-689.

[13] 林宏刚, 赵航宇, 陈麟. 基于语义增强的网络安全实体识别[J]. 计算机工程与设计, 2024, 45(9): 2584-2590.

[14] 高博, 常莉, 张金波, 等. 基于 BERT-CRF 和数据加密的石油勘探数据脱敏与安全保护研究[J]. 自动化与仪器仪表, 2025(3): 20-24.

[15] Zheng, X. and Wu, H. (2022) Autoregressive Linguistic Steganography Based on BERT and Consistency Coding. *Security and Communication Networks*, **2022**, Article ID: 9092785. https://doi.org/10.1155/2022/9092785

[16] Ueoka, H., Murawaki, Y. and Kurohashi, S. (2021) Frustratingly Easy Edit-Based Linguistic Steganography with a Masked Language Model. arxiv: 2104. 09833.

[17] Ding, C., Fu, Z., Yu, Q., Wang, F. and Chen, X. (2024) Joint Linguistic Steganography with BERT Masked Language Model and Graph Attention Network. *IEEE Transactions on Cognitive and Developmental Systems*, **16**, 772-781. https://doi.org/10.1109/tcds.2023.3296413