基于命名实体识别的数据库实体感知与安全查询优化

Azrin Sultana[a,*], Hasibur Rashid Chayon[a]

[a]孟加拉国美国国际大学计算机科学系，地址：孟加拉国达卡市 *B* 区 *4* 号路 *83/B* 号，邮编 *1229*

arXiv:2601.01254v1 [cs.DB] 3 Jan 2026

摘要

云存储已成为现代数据基础设施的支柱，但隐私保护和高效数据检索仍面临重大挑战。传统的隐私保护方法主要致力于提升数据库安全性，却未能在加密前自动识别敏感信息。这种自动识别可以显著缩短查询处理时间，减少人工识别敏感信息时出现的错误，从而降低潜在的隐私风险。为解决这一局限，本研究提出一种智能隐私保护查询优化框架，整合命名实体识别（NER）以检测查询中的敏感信息，并行采用安全数据加密和查询优化技术处理敏感数据与非敏感数据，从而实现高效的数据库优化。该方法结合深度学习算法与基于 Transformer 的模型，以高精确率检测并分类敏感实体；采用高级加密标准（AES）算法进行加密，并通过盲索引保障敏感数据的搜索功能；同时，使用 K 均值算法将非敏感数据划分为多个组，并通过排名搜索进行优化。在所有命名实体识别模型中，深度置信网络与长短期记忆网络相结合的模型（DBN-LSTM）表现最佳，准确率为 93%，精确率为 94%，召回率和 F1 分数均为 93%。此外，借助盲索引，加密搜索显著加快了搜索速度；非敏感数据的提取表现也优于传统的基于聚类的搜索。通过整合敏感数据检测、加密和查询优化，本研究推动了现代云基础设施中隐私保护计算的发展。

关键词：
命名实体识别、查询优化、数据库安全、高级

*Azrin Sultana. 电子邮件地址：azrin.sultana.cse@gmail.com

加密标准、盲索引、K 均值聚类、基于 Transformer 的模型

---

1. 引言

云计算使各类客户能够在线存储、分析和访问数据服务，大幅降低了计算成本 [1]。在云环境中，数据外包方式已变得极为普遍 [2]。因此，随着 NoSQL 数据库迅速发展（这类数据库能够存储海量非结构化和半结构化数据），越来越多的个人和企业希望将数据外包到云端 [3] [4]。然而，数据外包虽然使数据用户能够访问数据，却也不可避免地带来了数据泄露风险。不可信的服务提供商或攻击者可能窃取用户的个人信息，例如电子邮件、地址、电话号码和信用卡号，并将其出售给第三方。更重要的是，以云服务器为目标的攻击者能够获取客户的敏感个人信息 [5][6][7]。在所有数据泄露事件中，约有一半（46%）涉及客户的个人身份信息，例如家庭住址、电话号码、电子邮件和社会保障号码 [8]。根据 IBM 于 2023年的估算，云端存储的数据涉及 82% 的数据泄露事件。这类事件的成本高于平均水平，达 475 万美元；此外，39% 的数据泄露事件涉及多个环境 [9]。

数据加密通过防止云提供方获知其存储的数据，显著降低隐私泄露的风险。然而，手动识别和隔离此类个人数据，或重复使用基于规则的流程，尤其是在数据未标记或非结构化的情况下，并不具备可扩展性 [10][11]。因此，有必要依赖能够处理多样化且未标记数据的智能自动化技术，而这类数据在该领域中十分常见。将 NER 集成到该系统中，有助于自动高效地识别和保护敏感属性，从而在不影响性能的情况下实现安全的数据处理 [12][13]。NER 是一种至关重要的自然语言处理（NLP）技术，能够让机器识别并分类文本中的关键部分。NER 使机器能够将文本分类到指定类别中，从而帮助系统自动定位 NoSQL 查询中的敏感内容 [14][15]。自动提取敏感实体是实施以隐私为导向的方法（如加密、索引和访问控制）的第一步。如果没有 NER，要实现大规模隐私保护将是一项不可能完成的任务，这将导致对非敏感数据进行过度加密，或无法保护敏感数据，进而对安全性和查询性能产生负面影响。

在发生数据泄露或未经授权的访问时，银行记录、个人身份信息或医疗记录等加密信息对入侵者而言无法访问，因为他们无法获取密钥 [16]。然而，这种额外安全特性也存在一个令人担忧的

局限：无法通过常见查询方法访问加密数据 [17]。由于 NoSQL 数据库是无模式的，若不对密文进行处理，便难以对加密值执行搜索、排序和范围查询等常规查询 [18]。因此，加密会降低数据库的效率。于是，人们越来越需要能够直接处理加密数据的查询优化方法 [19]。此外，如果缺乏优化，对数据库进行查询可能会造成严重后果，尤其是在需要实时分析、快速数据传输和隐私感知系统的场景中[20]。在这些情况下，平衡加密与智能索引、聚类和排序机制，才能确保安全性不会损害系统的可用性。因此，开发隐私保护查询优化方法是现代 NoSQL系统的一项重要研究方向，特别是在组织寻求保护敏感数据，同时保持大规模查询性能快速且准确的情况下 [21][22]。

尽管大多数项目只专注于优化查询性能、增强数据库安全或识别隐私数据，但我们提出的模型的新颖之处在于它能够兼顾这两方面，通过自动识别和优化来保护用户数据隐私。根据我们的文献综述，现有研究尚未将命名实体识别用于识别敏感信息、在将其发送到数据库之前进行加密，并利用聚类来优化非敏感信息的搜索。处理数据库相关任务（例如加密信息和优化）需要深入了解数据库，并从系统层面理解安全数据流如何与智能信息检索共存 [23]。由于这些领域在常规研究或行业解决方案中鲜少交集，且需要大量专业知识，目前没有可用于支持此类工作的框架、工具或最佳实践。由于缺乏带有敏感数据标签的真实世界数据集 [24] ，这一领域的研究难以开展。因此，要训练、测试和基准评估旨在同时识别敏感数据并对其进行安全加密、同时优化查询的系统，也就十分困难。现有研究中的这些空白凸显了对一种框架的迫切需求：它不仅能通过自动检测来保护敏感数据，还能针对加密数据和明文数据实现智能查询优化。

在本研究中，为满足这一需求，我们提出了一种全面的隐私感知框架，通过命名实体识别模型、加密和安全索引技术来优化 NoSQL 数据库上的查询。首要目标是利用训练好的命名实体识别模型，自动识别敏感信息，包括电子邮件地址、电话号码和信用卡号码。识别出这些敏感实体后，便使用先进的加密算法对其进行加密。为支持对加密数据进行高效查询，我们采用盲索引。为进一步提升检索准确率和用户体验，我们采用基于词频-逆文档频率（TF-IDF）的向量化和 K 均值聚类，对非敏感信息的结果进行排序。

我们提出这项研究的目的，是在强隐私保证与高查询效率之间取得平衡，使其适用于金融、医疗保健和企业通信等隐私敏感领域。通过在安然数据集上进行实验验证，我们证明了所提方法在安全性、速度和检索相关性方面的有效性。

我们的贡献

- 一种由命名实体识别驱动的自动化流程，用于检测并分类非结构化 NoSQL 数据中的敏感数据（电子邮件、电话号码、信用卡号码）。

- 在此模型中，采用盲索引实现对加密数据的高效、安全查询，从而无需解密即可高效检索敏感数据。
- 集成了高级加密标准对称加密，以维护数据机密性和安全性，并通过加密文本安全地保护敏感数据。
- 在此模型中，采用盲索引实现对加密数据的高效、安全查询，从而无需解密即可高效检索敏感数据
- 为优化非敏感信息的搜索，系统加入了基于词频-逆文档频率的向量化系统和 K 均值聚类，以提高查询优化的相关性和针对性。

本文其余部分的结构如下：第二节介绍文献综述，第三节阐述所提出模型的方法论，其中我们将详细说明所提出的系统，并讨论优化技术。第六节介绍实验性能分析，第五节为讨论部分。最后，我们在结论部分给出本文的结论与未来研究方向。

2. 文献综述

数据外包和云存储的激增，彻底改变了当今数据管理领域对安全、高效数据检索的需求。为此，先前的研究提出了各种加密方案和索引技术，以实现对加密数据的安全快速搜索。然而，现有方法通常依赖人工识别敏感信息，或对整个数据库进行加密，这会引入人为错误并妨碍可扩展性。根据我们的模型，通过命名实体识别进行敏感信息检测，不仅能减少错误，还能实现更有效的加密和索引。以往的研究提出了独立的算法来优化加密方法；

然而，这些方法存在一些局限性。没有任何方法尝试将敏感数据识别、加密和数据库优化整合到一个综合框架中，而我们的模型正是针对这一点，旨在显著提升系统性能。

2.1. 识别命名实体识别

Dias 等人 [25] 提出了一种模型，能够从语料库中提取并分类葡萄牙语敏感数据。研究使用了标注了实体类别的葡萄牙语数据集、HAREM 数据集和 SIGARRA 新闻语料库。模型测试使用了 DataSense 命名实体识别语料库。命名实体识别基于三个子模块：用于识别敏感信息的基于规则的方法；用于识别社会经济信息的基于词典的分析；以及针对更模糊实体的机器学习模型，因为基于规则和基于词典的方法不足以处理这些实体。研究使用条件随机场、双向随机森林和长短期记忆网络训练并测试数据集，采用五折交叉验证和 80 个训练轮次以缓解过拟合；丢弃率为 0.68。模型的 F1 分数为 83.01%。

Sun 等人 [26] 利用多种深度学习模型，提出了一种用于自然灾害命名实体识别的预测模型，并根据三个步骤分析这些模型，包括预训练、特征提取和解码。该模型采用了万方数据库中的研究论文，共收集 12,387 篇仅包含标题和摘要的论文。四位作者共同完成了数据集标注。本文使用来自变换器的双向编码器表示（BERT）、ALBERT 和 XLNet，在大规模未标注的自然灾害语料库上进行训练。研究使用开始、内部和外部（BOI）标签标注实体，而 ALBERT 用于句子顺序预测。所提出的模型使用 XLNET 准确识别上下文词语。研究使用双向长短期记忆网络和双向门控循环单元提取重要特征。通过结合 XLNET、双向长短期记忆网络和条件随机场（CRF），该模型取得了最佳性能，精确率为 92.80%，召回率为 91.74%，F1分数为 92.27%。

为识别非结构化文档中的敏感信息，Ahmed 等人 [27]采用了深度学习、机器学习和基于规则的方法。研究使用了两个公开数据集：其中一个是从 Kaggle 收集的文本敏感型上下文相关数据集，包含 100 万个实例；由于缺少标签，因此选取了 400 个实例。另一方面，图像数据集来自三个来源，包含 300 张标记为敏感或非敏感的图像。他们使用留一交叉验证对模型进行了验证。研究使用了三种不同类型的模型，即基于机器学习、基于规则和基于深度学习的模型，并比较了它们的结果。为验证图像中是否缺少字母或数字，研究使用了莱文斯坦距离度量

来衡量词语与关键词之间的距离。在这三种模型中，基于深度学习的模型表现最佳，准确率达到 95%。

为从非结构化文本中提取患者报告数据， [28] 采用了一个弱监督模型，该模型使用医学命名实体识别模型在八种语言上进行训练。第一步是合并多个医学文本数据库，构建完整的语料库。为缓解类别不平衡问题，研究人员对数据集进行了数据增强，这也有助于缓解过拟合。微调后的模型在英语上的表现最佳，F1 分数为 80.07%；其次是德语、西班牙语和葡萄牙语，这三种语言的得分均超过 77%。在资源较少的语言中，斯洛伐克语、意大利语和波兰语的结果颇具潜力，得分均为 75%。

论文 [29] 采用了特征提取、视觉特征、文本特征和多模态特征融合这五个组成部分，提出了一套高质量方案，用于在电力设备隐私敏感数据集中构建隐私数据。数据通过网络抓取获取，并采用 BIO标注方案进行标注。该模型由三个主要部分组成：文本特征提取、字符嵌入构建和 CRF维特比解码器。与基线方法相比，采用多头注意力的基于BiLSTM的特征提取取得了更好的性能。

通过对抗训练和外部上下文检索， [30] 提出了一种用于识别中文汽车缺陷消息的命名实体识别模型。该模型利用搜索引擎检索字符—词融合嵌入表示，并采用改进的关键词提取方法，生成与原始输入短语在语义上相关的外部文本。随后，使用注意力机制融合字符—词融合嵌入表示。最后，使用双向长短期记忆网络提取长距离语义信息，并使用条件随机场获取识别结果。在领域专用数据集和通用命名实体识别数据集（如微博、简历）上进行的实验表明，该模型优于现有模型，在汽车数据集上取得了89.41%的F1 分数，显示出较强的泛化能力和鲁棒性。

更多与命名实体识别检测综述相关的文章见表1。

表1：命名实体识别研究综述

发表年份 Year	**目标**	**数据集**	**算法**
Wilkho et al. 2025 [31]	开发 a 中文新闻文本的命名实体识别方法 文本	新闻领域中文命名实体识别语料库	迁移学习（ERNIE）和软词典嵌入

发布 Year	目标	数据集	算法
Gao等人等，2025年 [32]	开发一个旅游-专用 NER 模型 that 预测实体边界，以有效识别嵌套的和长实体旅游领域中的	三个旅游数据集结合公共和私人来源	双向长短期记忆网络
H. Cho et al. 2019 [33]	提取信息 from 生物医学文章	NCBI 疾病语料库, BioCreative II 基因提及，化学物质-疾病关系语料库	双向长短期记忆网络, CRF, BERT, 卷积长短期记忆网络
Zhou 等人（2025）[34]	基于深度学习的 NER for 提取信息从建筑施工中事件报告	建筑施工现场事件报-告	BERT, BERT-LSTM-条件随机场、BERT-LSTM
Serere et al. 2025 [35]	动态命名实体识别模型检测作者立场和地点指代	Twitter 语料库 of 4,098 条帖子（哈佛数据宇宙）	LLM
G. et 等，2025 [36]	用于区分当前与远程位置的动态命名实体识别网络安全推文中的网络安全推文	自定义 Twitter 数据集已标注使用语法规则	微调后的大语言模型
Wu et al. 2018 [37]	提取概念性临床文本	MIMIC II 语料库	循环神经网络、卷积神经网络
Zhang et al. 2025 [38]	兼顾安全与隐私的感知型命名实体识别，用于实体提取	Data from Whois, Censys 和 Shodan	BERT, GLM4-9B, Mixtral-8B、LLaMa3-8B 模型
Wang 等人等人，2024年 [39]	少样本多视图用于网络安全的命名实体识别威胁情报	威胁情报数据集	BERT, BiL-STM+条件随机场

发布 Year	目标	数据集	算法
Demirol et al. 2025 [40]	提取网络威胁实体（行为者，malware、攻击活动、目标）	100 份 PDF报告 + 540 HTML 页，人工标注	BERT

2.2. 保护并优化数据库

为优化数据库操作，加密数据必须支持搜索和排序，这意味着必须暴露某些信息。为此，Wiese等人 [41]开发了一个框架，可在 NoSQL 数据库中未经修改的宽列存储上实现保属性加密的最高安全性、功能性和性能。他们采用基于索引的加密方案来扩展数据库。为了在加密数据上执行模式，他们使用 Java密码扩展密钥库存储通过自定义标签检索的一组随机数量的密钥。可搜索加密仅针对字符串执行，而保序加密则适用于字符串或数值，这意味着每列的索引都得以保留。所提出的模型测量了加密查询与检索相同结果之间的通信时间。

Du等人 [42] 将客户端的授权数据纳入搜索令牌和索引中，并提出了一种多客户端对称加密方案。该模型使多个客户端能够在加密数据库上执行布尔查询，同时将搜索能力限制在特定关键词范围内，从而提供双层安全保护。当恶意用户使用其授权范围之外的关键词进行查询，并试图访问未获授权的内容时，系统会阻止该请求。该模型根据计算和通信成本进行了评估。它支持涉及布尔运算（例如 AND、OR、NOT）的复杂搜索查询，从而能够表达更丰富的搜索条件。

He等人 [43] 提出了一种支持排序搜索的快速可搜索加密方案，并采用聚类算法对文档进行聚类。聚类时，使用关键词转换方法将文档和查询转换为向量。随后应用树构建算法，为每个文档集生成一棵索引树，并计算查询与索引树根节点之间的相关性得分，再使用对称加密选出最高相关性得分。对比结果表明，查询效率随文档规模增长呈次线性变化。

Elizabeth等人提出了一种针对加密云数据的安全排序搜索方案[44]。该方案为关键词构建加密可搜索索引，并为搜索关键词构建相应的陷门。利用加密得分索引，该方案验证并计算查询关键词的得分，同时对文档中的相关查询进行排序，并提供可验证参数。对于陷门和索引的生成，

该方案采用 Paillier密码体制，确保第三方和提供方均无法获知排序信息。该模型可动态更新加密字段，而无需重建整个索引。为验证查询结果，该方案利用带密钥哈希实现可验证搜索。在安然数据集上进行的模拟表明，所提出的方案在语义上安全，并能保障正确性和隐私。作者为验证生成输出和更新查询场景做出了重要贡献。然而，系统越复杂，所需空间就越大；文中未讨论如何处理空间复杂度。此外，文中未提及哪些信息被视为敏感信息，且该方案仅针对单个数据所有者设计，也未说明该系统的工作方式。

本文提出了一种多关键词可搜索加密方案 [45] ，用于实现安全且高效的搜索，尤其是在不损害安全性的前提下支持多关键词查询。该方案包含五个阶段：密钥生成、索引构建、陷阱生成、查询处理和结果检索。其核心创新在于使用基于概率的陷阱和关键词向量，以增强搜索安全性和查询灵活性。该方案能够抵御基于概率的陷阱构造，从而保障其安全性。作者证明，这种概率设计可以防止搜索模式被逆向推断以及关键词泄露。

[46] 探讨了云环境中多用户、多关键词场景下，对安全且保护隐私的可搜索加密日益增长的需求，并采用双服务器模型和私有子集判定机制。该方案确保数据隐私和查询隐私，实现定长密文和陷阱，并支持云环境中的多用户访问。该方案采用密码学原语进行安全的比特分解，并使用陷阱密钥密码体制支持安全搜索操作。此外，文中还分析了基于模拟的安全性，确保在互不串通的半诚实实体面前仍具有强隐私保护能力。实验结果表明，与以往方法相比，所提出系统的陷阱生成、密文大小和查询效率均更优，尤其是在大规模场景下。

许多研究采用从基于规则的系统到深度学习方法等各种方法来处理敏感信息检测。虽然在命名实体识别和 Transformer模型的辅助下，这些方法效果良好，但泛化能力不足，且计算开销大。另一方面，面向加密的模型通过改进索引来支持搜索，却未能充分解决隐私可扩展性和粒度方面的问题。与此同时，其他研究者提出了安全、排序和可验证搜索模型；然而，这些模型受限于复杂性、单用户架构，以及对敏感数据定义不一致等问题。尽管如此，目前还没有研究将敏感信息自动检测与 NoSQL 数据库优化结合起来。为克服上述缺陷，并弥合数据库优化与利用命名实体识别模型自动识别敏感信息之间的差距，我们提出了一种高效的命名实体识别模型，其计算开销低，因为我们没有使用任何高

级模型。然而，我们的模型并未牺牲性能。我们只对敏感信息进行加密，而非加密整个数据库，这是效率最高、速度最快的方法。为保障数据库安全，我们通过实施双层安全，并对非敏感数据进行索引和聚类，以实现高效检索，从而对数据库进行了优化。

3. 方法论

本章介绍了为设计和实验所采用的流程，旨在构建所提出的隐私保护型数据识别与数据库框架优化方案。该方法集成了命名实体识别模型，用于自动识别敏感信息，随后通过加密和索引实现安全的查询处理。此外，还对非敏感数据执行聚类、向量化形式搜索和排序搜索，以提高检索效率。本章详细解释了系统的各个阶段，包括预处理步骤、加密步骤以及用于评估的性能指标。针对本研究，所提出的系统将命名实体识别与数据库安全和查询优化相结合，这是该领域的一项基准方案。利用命名实体识别自动识别敏感信息，例如联系电话、电子邮件地址、社会保障号码和信用卡号码，可以免除手动处理敏感信息所需的额外工作。除所提出的方法外，该系统还在不牺牲安全性的情况下实现了高效的查询优化。

3.1. 数据集描述

本研究所用的数据集采集自 Mendeley [47], ，其中包含经过匿名化处理的个人身份信息。该数据集模拟了金融场景中常见的各种个人实体，旨在支持敏感数据的开发与评估。每条记录都模拟真实世界的信息。数据集包含 45,000 个实例和 10 个属性，分别为姓名、社会保障号码、信用卡号码、电话号码、电子邮件地址、实际地址、公司名称、文本和网址。本研究仅使用了文本属性。数据集的大多数文本属性中都包含敏感信息。我们谨慎收集并细致标注了这些数据，以便准确识别其中的敏感信息。随后，这些已识别的实体将用于检测 NoSQL 数据库中的敏感信息。表2列出了从该数据集中收集的敏感信息详情。文本平均长度为 1,357 个字符，内容全面详尽，有助于清晰理解并准确获取信息。为验证我们的数据库优化模型，我们使用了安然数据集，其中包含电子邮件地址等敏感信息，也包含非敏感信息，因此适合用于验证查询优化系统。

表2：数据集中的实体详情

类别	实体名称	敏感数据包含
个人身份认证号码	护照	护照号码
	电话	电话号码
	信用 _卡	信用卡号
	电子邮件	电子邮件地址
社会经济信息	SSN	社会保障号码-ber
其他	URL	Link
	支付	货币

3.2. 模型描述

我们提出的研究分为三个阶段。第一阶段，用户提交查询后，我们通过命名实体识别模型，利用深度学习和自然语言处理识别电话号码、电子邮件地址、社会保障号码、链接和信用卡号等敏感信息，然后使用加密技术对已识别的敏感数据进行加密。由于数据库属性加密不同于常规加密，这构成第二阶段，即隐私保护阶段。我们必须确保加密值在内容仍被隐藏的同时，仍可被完整搜索并支持条件查询，从而实现高效且安全的数据利用。所提模型的最后阶段是应用数据库优化技术来提升性能、速度和安全性，即使敏感数据经过加密并存储在数据库中也是如此。

3.2.1. 命名实体识别模型

命名实体识别模型用于从非结构化数据中识别敏感信息。这是我们所提出模型的第一步。如图1所示，数据集依次经过分割、分词、标点符号去除、定义正则表达式（regex）、IOB格式标注和向量化等预处理步骤。预处理完成后，使用深度学习模型对该模型进行训练、测试和验证。每个步骤将在下文中详细讨论。

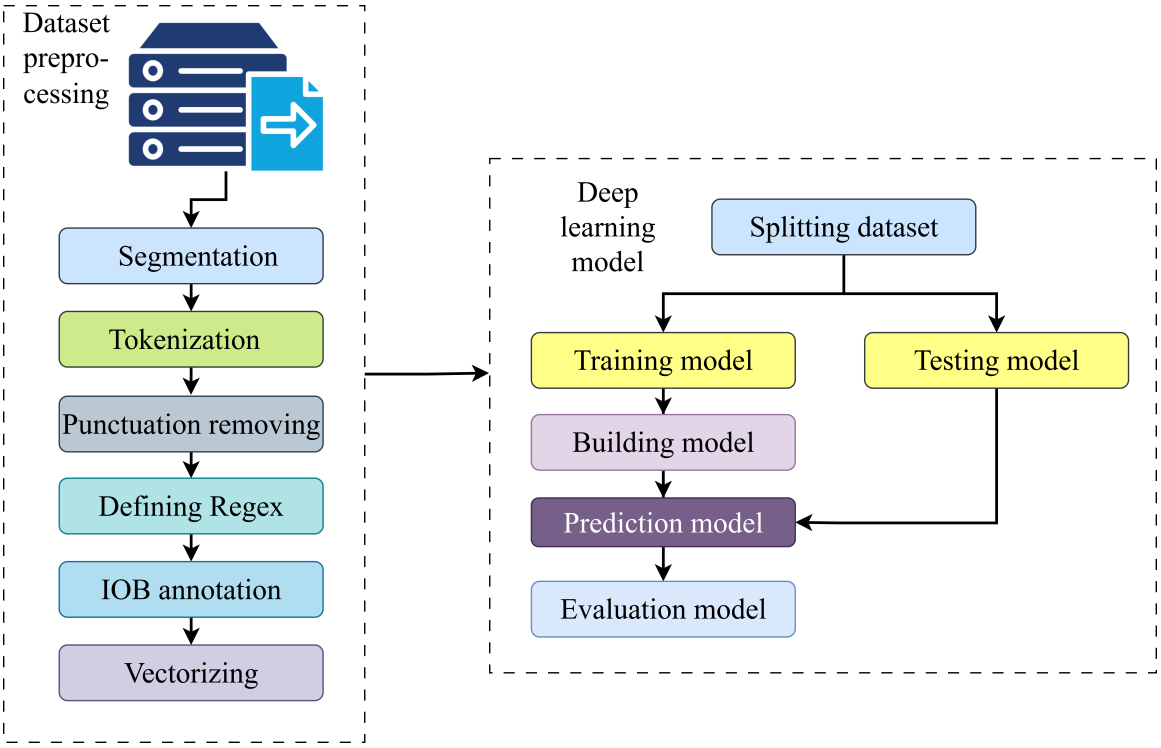


图1：命名实体识别模型流程图

表3：文本的词元化

句子	词语
致 Andrade有限责任公司的成员	'To', 'the' , '成员' , 'of', '安德拉德', '有限责任公司'
年度合规之后	'之后', '的', '年度', '合规-ance'
独立审计师关于.....的报告	'独立', '审计师", '报告', '关于'
我们进行了全面的检索。	'我们', '已经', '进行了', '一次', '全-面', '检索'
敬启者，关于此事,	'敬启者', '关于', '此事', ', ' '此事'

3.2.1.1 分词

分词是指将原始文本拆分成更小的词语或句子片段，这些片段称为词元。将句子拆分成词语称为词语分词，而句子分词是指将文本拆分成独立的句子。进行词语分词时，主要依据词语之间的空格；而句子分词则使用感叹号和句号等字符 [48]。在我们提出的模型中，采用词语分词将句子拆分成独立的词语，如表3所示。

3.2.1.2 从文本中删除标点符号

从文本中删除标点符号是自然语言处理中的关键步骤。标点符号，即句点、逗号、冒号和问号，本身并无显著意义；然而，它们有时会使结果产生偏差，并且由于词频和嵌入层的影响，模型可能会在噪声数据上进行训练 [49]。因此，删除标点符号可以简化文本，通过降低数据中的复杂性和不一致性来调整分析过程 [50]。在此模型中，文本预处理阶段会删除标点符号。删除所有标点符号会影响我们的数据，因为许多敏感词中都包含标点符号，例如句号、连字符和逗号。电子邮件地址中含有句点，而付款金额中可能会用到逗号。此外，社会安全号码、护照和信用卡信息中也可能包含连字符。因此，我们设计了一种算法，只删除词语末尾的标点符号，以免影响我们需要的值。表4列出了我们删除标点符号的详细信息。

表4：移除标点符号后的清理后的词语

词语	清理后的词语
jonathon@gmail.com,	jonathon@gmail.com
10,500. 0987345567689,	10,500 0987345567689
(017123455688)	017123455688

表5：用于识别数据集敏感信息的正则表达式

Name	实体	正则表达式
电子邮件地址	电子邮件	[a-zA-Z0-9.%-]+@[a-zA-Z0-9.-]+ [a-zA-Z]2,+
电话号码	电话	\+(\d+)(?:\s?(\d{6,}))?
社交 安全号码	SSN	\b\d{3}-\d{2}-\d{4}\b
付款金额	货币	\$\d{1,3}(?:,\d{3})*(?:\.\d{2})?
信用卡号	信用 _卡	\b\d{13,16}\b
Link	URL	https?://[^\s]+

3.2.1.3 定义正则表达式

正则表达式（Regex）是用于搜索模式的一种词语排列方式 [51]。我们使用正则表达式从文本中识别敏感信息。首先，我们仔细观察了文本中敏感信息的模式，并据此设计了用于识别这些信息的正则表达式（Regex）。在删除文本中的标点符号后，我们应用这些正则表达式。表5介绍了用于确定各项值的正则表达式。

3.2.1.4 IOB格式数据集标注



图 2：IOB格式标注

数据集标注是将数据标记为预定义类别的方法。这是一个至关重要的过程，尤其适用于监督学习，并且可用于多种类型的数据集。数据标注在评估模型性能方面发挥着关键作用，因为它能够提供目标标签。至少应由两名人员共同补充标注过程，以验证标注结果的可信度 [52] [53]。语料库的标注有多种技术。IOB指标签的内部（Inside）、外部（Outside）和开始（Beginning），这些部分共同构成IOB。图2展示了IOB格式。随着LSTM、卷积神经网络和Transformer模型在序列标注中对IOB的理解越来越准确，它们能够以更高的准确率训练模型。此外，当数据集实体采用IOB格式时，使用性能指标更容易验证所提出的模型 [54]。因此，我们采用该技术对数据集进行标注。在基于正则表达式完成标注后，我们手动验证了标注的准确率。如果发现文本标注中存在任何处理不当或错误，我们会解决该问题，修改算法标注文本的逻辑，然后重新运行算法1进行标注。数据集标注的流程图如图3所示，展示了使用IOB格式进行完整标注的过程。首先对数据集进行预处理，然后应用算法识别敏感文本，并使用IOB方案对其进行格式化。最后，我们手动检查是否存在错误识别，并据此重写算法1。

3.2.1.5 词向量转换

机器学习算法无法理解文本。因此，我们需要将文本转换为数值形式。词嵌入是一种用于以相应数值表示词语和文档的方法，这些数值被称为词向量。词向量在低维空间中描述一个词语，使含义相近的词语能够具有相似的表示。通过这种方法，可以从文本中提取特征，然后将其输入机器学习模型，同时保留句法和语义信息 [55] [56]。Word2Vec 是一种用于生成词嵌入的神经网络方法。在自然语言处理中，这是一种广泛使用的技术，用于将词语表示为连续向量空间；该空间通过将词语映射到高维向量来收集词语之间的语义关系，并为每个词语分配一个向量 [57]。在本研究中，我们使用 Word2Vec 将词语转换为向量。与其他传统方法相比，它能够生成彼此距离更近的稠密语义词组，这有助于降低复杂性并缩短训练时间。较低的复杂性、更少的训练需求以及对稠密向量的有效表示，是用于敏感信息识别的数据集分类最合适的特征

算法1 命名实体识别的词元标注

```
过程标注_词元(tokens, patterns)
    标签 ← 长度为 |词元| 的"O"列表
    活动_实体 ← 无
    对于 i = 0 到 |词元|-1 执行
        词元 ← 词元[i]
        实体 _found ← false
        对于实体，模式在模式集合执行
            如果 词元匹配模式 则
                如果 active_实体 ≠ 实体 则
                    标签[i] ← "B-" +实体
                    活跃_实体 ← 实体
                else
                    标签[i] ← "I-" +实体
                结束 if
                实体 _已找到 ← 真
                分隔符
            结束如果
        结束循环
        如果 不是实体 _已找到则
            活动_实体 ← 无
        结束 if
    结束 for
    返回 标签
结束过程
```

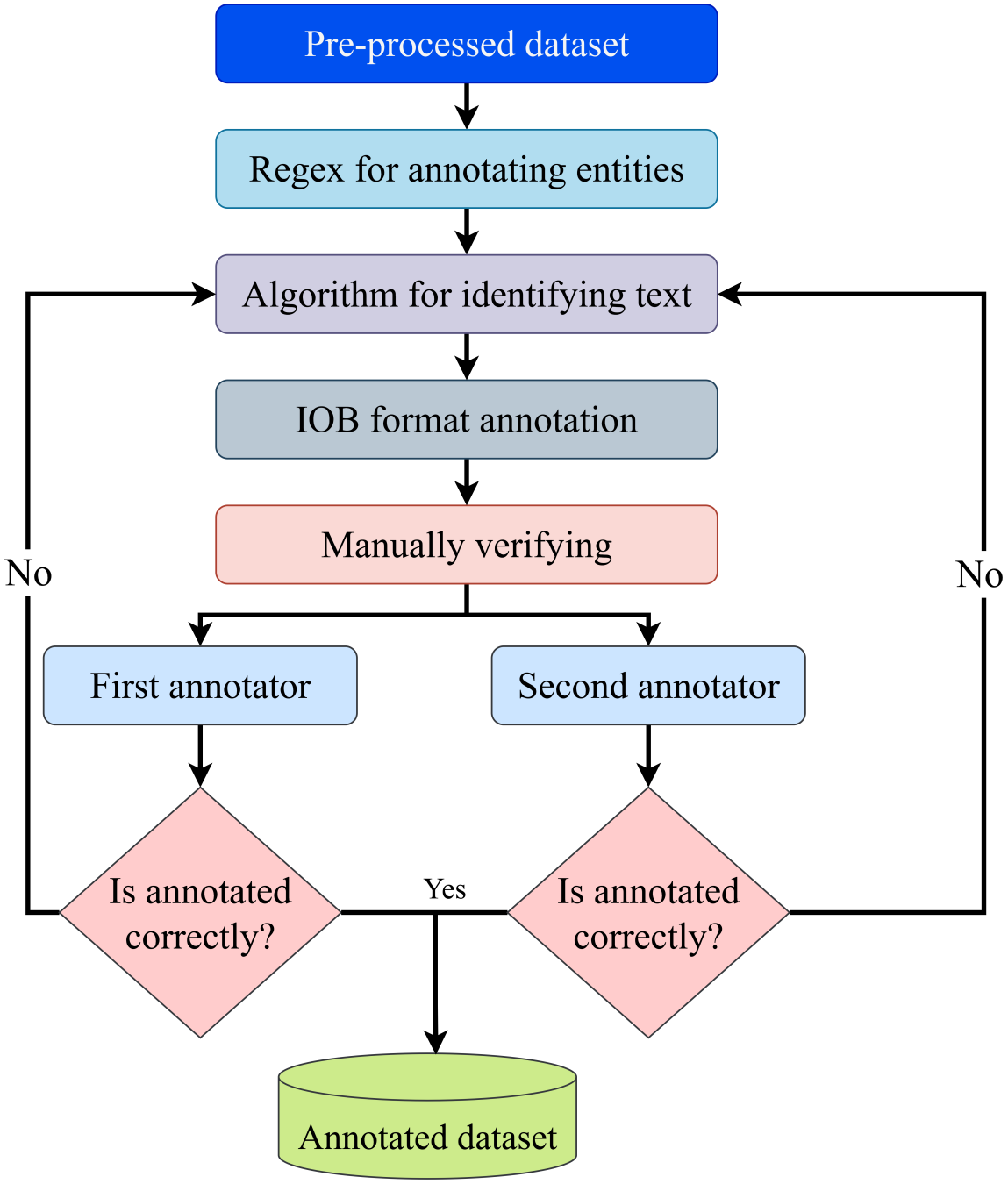


图3：IOB格式标注流程图

3.2.1.6 训练命名实体识别模型

完成预处理并将文本转换为机器可理解形式后，我们按照 80:20 的比例对其进行划分，并将其输入深度学习模型。我们使用深度学习算法以及混合模型对模型进行训练。我们采用了三种深度学习算法并构建了混合模型，具体内容将在算法部分详细讨论。训练完成后，我们对模型进行了测试，性能详情见结果部分。

3.2.2. 确保数据库安全并进行优化

命名实体识别模型对数据进行分类后，系统会生成密钥并使用盲索引对敏感数据进行加密，以便高效检索。相比之下，非敏感数据会被向量化、聚类，并根据余弦相似度分数进行排序，从而加快搜索速度。该流程的示意图见图4。

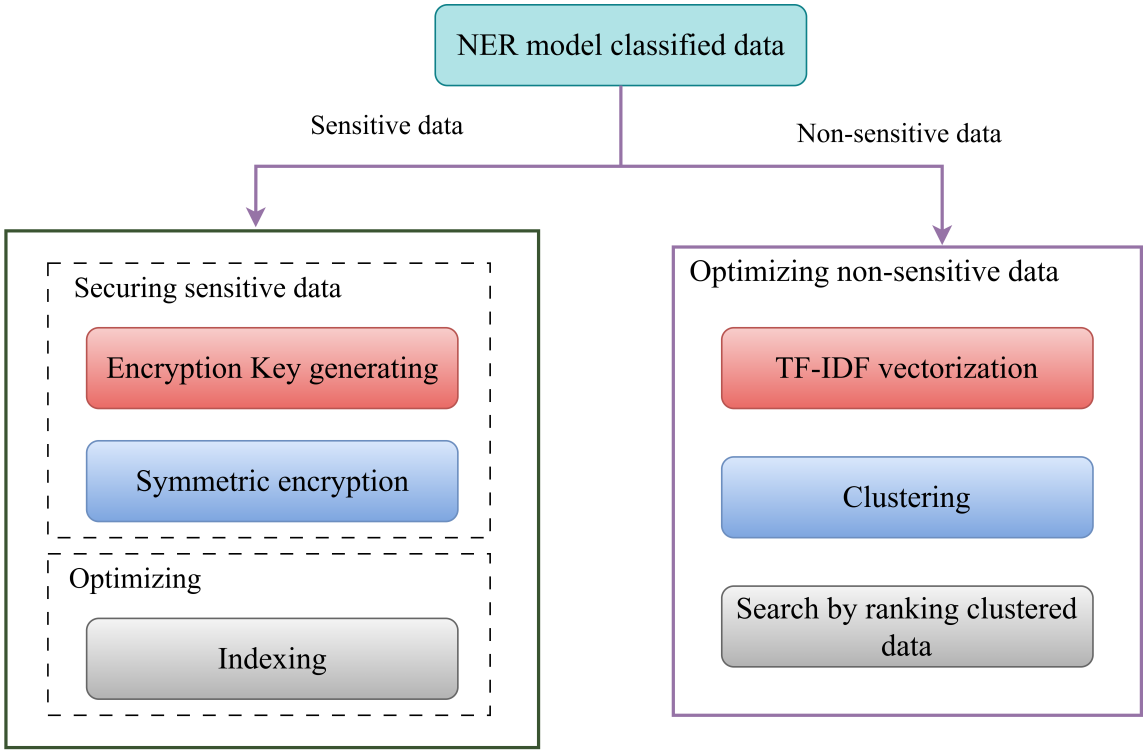


图4：数据库安全与优化流程图

3.2.2.1 通过生成密钥加密数据

对称加密使用同一个密钥进行加密和解密，是一种广泛应用的密码学形式，能够提供速度更快、更易部署的安全保障。它主要用于数据处理，其中效率、速度和复杂性是关键因素 [58]。高级加密标准是应用最广泛的对称加密算法之一。高级加密标准有多种不同的复杂度，例如128位和256位，代表着逐步提高的安全级别，并能有效防范未授权访问。它效率高，通常用于保护敏感数据 [59]。通过命名实体识别模型识别出敏感信息后，我们会使用AES算法对其进行加密。由于几乎不可能破解，它能够确保敏感数据的机密性。

3.2.2.2 敏感数据索引

盲索引是一种索引技术，可在不泄露实际数据的情况下搜索加密数据。它不会直接为敏感列建立索引，而是使用确定性密码哈希函数（通常是带密钥的 HMAC）生成明文的定长表示。该哈希值称为盲索引，并与加密数据一同存储 [60]。执行搜索时，客户端使用相同的函数和密钥对搜索词进行哈希处理，生成相应的盲索引值，再将其用于数据库查询。由于相同的输入始终会生成相同的哈希值（使用同一密钥时），因此即使服务器从未接触明文，也能支持精确匹配搜索。盲索引在隐私保护场景中特别有用，因为电子邮件、电话号码或信用卡号码等字段需要加密，同时仍须支持搜索。它能优化敏感信息的处理，在不损害数据隐私的前提下实现高效、快速的查询。因此，盲索引通过使用安全且确定性的哈希（盲索引），直接对加密数据进行精确匹配搜索，从而减少通过解密所有记录来执行搜索所带来的开销，并显著提升查询性能。即使面对海量加密数据，它也能让数据库支持低延迟搜索。因此，盲索引不仅增强了数据隐私，还通过结合密码学安全性与数据库索引优化来提升性能 [61]。

3.2.3. 非敏感数据优化**3.2.3.1 词频-逆文档频率（TF-IDF）向量化**

该算法基于这样一个原理：衡量一个词在某篇文本中的重要性，并将其与一组文本进行比较。它是一种广泛用于文本处理的文本嵌入方法。其中，TF 表示词频，用于确定某个词在文本中出现的频率 [62]。其计算方法是将某个词的出现次数除以文档中的值总数。此外，IDF 表示逆文档频率，用于衡量某个词在语料库所有文本中的稀有程度。它通过对数函数计算。因此，词频-逆文档频率（TF-IDF）是 TF 和 IDF 的乘积，会为在某篇文本中频繁出现、但在整个文档语料库中较为少见的词赋予较高的值。此外，IDF 表示逆文档频率，用于衡量某个词在语料库所有文本中的稀有程度。它通过对数函数计算。因此，TF-IDF 是 TF 与 IDF 的乘积，会为在某篇文本中频繁出现、但在整个语料库中较为少见的词赋予较高的值[63]。这里使用这种向量化技术，为与上下文相关的词赋予较高权重。

3.2.3.2 非敏感信息聚类

聚类是对相似类型的数据进行分组的最常用技术之一，即识别数据的子组，使同一簇内的数据点彼此高度相似。相反，不同簇中的数据点具有高度差异，从而为多样化数据提供更高的价值 [64] [65]。K均值算法是一种无监督算法，将数据集划分为 N 个预定义的不同子组。在分配数据点时，应确保簇的质心以及数据点之间的平方距离最小。当簇中的变化较少时，数据点会变得更加标准化 [66]。对于我们提出的研究，基于相似性对数据库进行聚类，对于优化和保护隐私都至关重要。我们的系统将基于由词频-逆文档频率生成的相似度向量创建簇。图5展示了数据库的聚类和排序。当授权用户搜索结果时，系统不会搜索整个文档，而是首先查找簇，因为每个簇都已进行排序。随后，系统会在该簇内搜索目标值。在数据库包含 200,000 条记录的情况下，检索某个值需要搜索整个数据库，这既耗时又低效。然而，根据相似性，如果某个簇仅包含 1000 条记录，那么要获得所需结果，只需搜索 1000 个数据点。

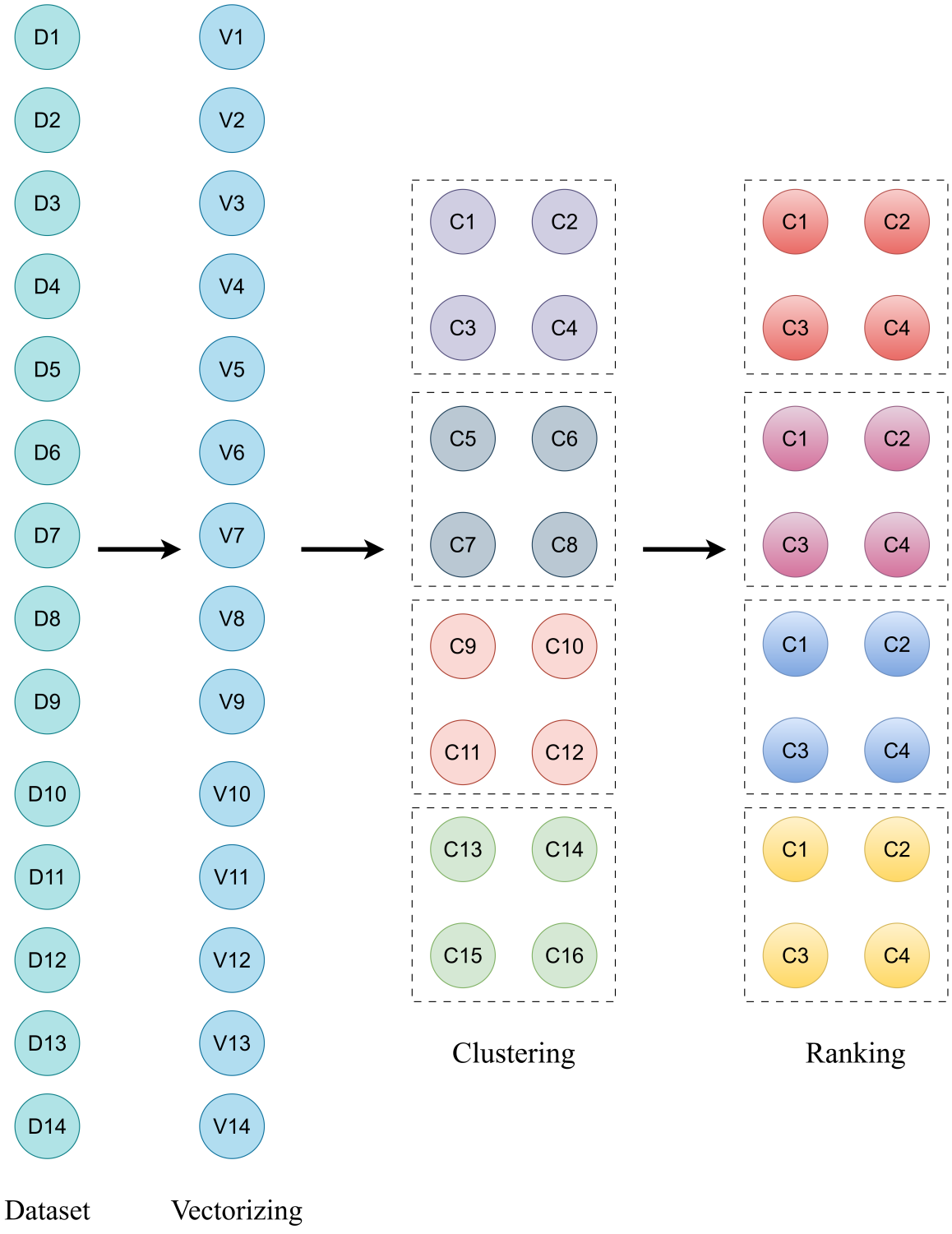


图5：对数据库中的非敏感信息进行聚类和排序

3.2.3.3 基于余弦相似度分数的聚类数据排序

余弦相似度通过测量两个非零向量之间夹角的余弦值来计算它们之间的相似度。在机器学习和数据分析中，尤其是在文本分析、搜索查询、文档比较和推荐系统中，余弦相似度得到了广泛应用。该分数通过根据数据集中特征的维度比较数据对象之间的距离来计算。距离越小，表示相似度越高；距离越大，则表示相似度越低。它是一种用于确定数据对象相似程度的度量，与数据对象的大小无关。在数据集中，测量簇中句子之间的相似度被视为一个向量 [67]。对于查询向量 Qv 和文档向量 Dv，余弦相似度分数为（1）：

$$\text{CosineSimilarity}(Q_v, D_v) = \frac{Q_v \cdot D_v}{\|Q_v\| \times \|D_v\|} \tag{1}$$

其中， $Qv.Dv$：是查询向量与文档向量的点积， $|(|Qv|)|\times||Dv||$是向量的欧几里得范数。

3.3. 敏感信息检测算法

3.3.1. 双向长短期记忆网络 （双向长短期记忆网络）

双向长短期记忆网络是一种强大的神经网络架构，它基于长短期记忆网络的能力构建而成。长短期记忆网络是一种专为序列建模设计的循环神经网络[68]。它由两个长短期记忆网络层组成：一个从头到尾处理输入（前向方向），另一个从尾到头处理输入（后向方向）。两者的输出相加，使模型能够同时考虑每个词的历史和未来，从而更好地理解其含义。每个时间步的输出会拼接起来，生成最终输出y_t：

$$\vec{h}_t = f(\vec{W}_x\vec{x}_t + \vec{W}_h\vec{h}_{t-1} + \vec{b}_h) \tag{2}$$

$$\overleftarrow{h_t} = f(\overleftarrow{W}_x x_t + \overleftarrow{W}_h \overleftarrow{h}_{t+1} + \overleftarrow{b}_h) \tag{3}$$

$$y_t = [\vec{h}_t, \overleftarrow{h}_t] \tag{4}$$

此处， $\vec{h}_t$和 $\vec{h}_{t-1}$是前向方向的状态向量。 $\overleftarrow{h_t}$和 $\overleftarrow{h}_{t+1}$是后向方向的状态向量。 $\vec{W}_x$和 $\overleftarrow{W}_x$分别是前向和后向方向的输入矩阵。 $\vec{W}_h$和 $\overleftarrow{W}_h$分别是前向和后向方向的权重矩阵。最后， $\vec{b}_h$和 $\overleftarrow{b}_h$是偏置。

双向长短期记忆网络对于命名实体识别（NER）等任务非常有用，因为理解词语周围的上下文能够提高实体识别的精确率。双向长短期记忆网络对查询理解和安全数据处理有重要贡献。它通过全面扫描单个词语的上下文，能够有效识别多词命名实体，例如医生姓名、组织名称和时间戳。此类双向处理再次支持对语义模糊或领域特定词的进一步理解 [69]。

3.3.2. 卷积神经网络（*CNN*）

深度学习算法（例如卷积神经网络）采用称为卷积核的卷积滤波器，对输入的词嵌入进行扫描，以寻找有意义的局部模式，例如前缀、后缀或固定表达式。这些滤波器能够从小段文本中提取高级特征，同时池化层对维度进行下采样，从而支持高效计算 [70]。卷积神经网络尤其适用于图像分类、目标检测和分割等任务，因为卷积神经网络能够通过反向传播自动学习并自适应地发现特征空间层次结构[71]。它主要依靠三种类型的层来高效学习空间和时间交互关系：卷积层是卷积神经网络的核心。每个滤波器都经过设计，用于通过在小型空间区域内计算逐元素求和与乘法来检测局部特征，例如边缘、纹理或形状，同时覆盖整个输入的深度[72] [73]。

3.3.3. 深度信念网络

深度信念网络是通过堆叠多个连续的受限玻尔兹曼机层构建的高阶概率生成模型。它们尤其擅长进行层次化特征学习，能够从原始无标签数据中逐步发现越来越高层次且更加抽象的表示。深度信念网络通常分两个阶段进行训练：第一阶段是无监督预训练阶段，在此阶段中，每个受限玻尔兹曼机层都以贪婪方式进行学习，以便预先最优地初始化网络权重；第二阶段是有监督微调阶段，用于针对特定任务对模型进行微调。这种两步训练能够使深度信念网络成功地从噪声或不确定输入中解读出实际特征。在序列模型中，深度信念网络通过促进形成更丰富、更深层的表示，同时保留输入数据的自然结构，从而在命名实体识别等任务上展现出更优异的性能，使模型能够更有效地处理语义和上下文 [74] [75]。

3.3.4. 双向编码器表示自变换器（*BERT*）

BERT 是较新的基于变换器的语言模型之一，旨在通过从两个方向同时分析每个词的上下文——左侧和右侧——来学习深层的词义知识

。它的全局视角使模型能够更好地捕捉更细微的语言现象和远距离词依赖[76]。第一种是掩码语言建模，即随机遮盖句子中的词语，并提示模型仅根据未被遮盖的词语所提供的上下文来预测这些词语 [77]。第二种是下一句预测，旨在训练模型预测给定的两个句子在顺序上是否具有逻辑上的先后关系，从而学习句子之间的语义关系[78]。

3.3.5. XLNET

XLNet 是一种基于 Transformer 的模型，可按从左到右或从右到左的顺序处理文本。该方法无需依赖掩码标记，也无需人为屏蔽部分词元，即可高效处理双向上下文 [79]。XLNet 采用 Transformer-XL架构，其特点是通过递归机制将上下文窗口扩展到固定长度的片段之外，从而高效处理长距离文本依赖。因此，XLNet 特别适合学习和处理较长文本，因为它能够捕捉相距较远的词语之间的关系。通过按随机顺序预测词元，XLNet 能够学习更高层次依赖关系的模式，并更有效地进行语言建模。这使其在各种自然语言处理任务中表现更佳，包括语言建模、文本分类和实体识别。XLNet 预测词元 [80]

3.4. 所提模型的隐私分析

根据所提出的模型，敏感信息通过加密和盲索引来确保隐私，而非敏感信息则使用安全向量化搜索。敏感信息的盲索引是一种技术，可在不直接泄露任何敏感信息的情况下，搜索加密或哈希数据。当用户插入包含敏感信息的数据（例如电子邮件、电话号码或信用卡号）时，我们的命名实体识别系统会通过提供查询 Q 自动识别这些信息，并识别后对其进行加密。对于此类数据的加密，我们使用了 AES 算法。当数据所有者（DO）向云服务器（CS）插入一条记录时，会向 CS 发送查询 Q。系统（S）接收到 Q 后，利用我们的命名实体识别模型识别出所有不能向 CS 泄露的敏感信息，并使用 K 生成密钥；S 对 Q 中的私人信息进行加密。使用相同的 K，S 为该值生成盲索引。在 CS 中，索引和 Q 均会存储。因此，除了被视为可信的 S 之外，CS 不掌握有关 Q 和 K 的任何信息。即使有人侵入 CS，也无法反向还原 Q 并获取其中的信息。对于非敏感部分，我们将文本转换为向量化形式。因此，即使不加密普通文本，CS 也无法收集

信息，因为所有非敏感信息都通过词频-逆文档频率（TF-IDF）转换为数字形式；捕捉其语义内容和词项重要性起初是一项艰巨的任务。

3.5. 查询优化

由于机密性与查询效率之间需要权衡，加密数据通常会增加搜索的复杂度。因此，为了提高效率，我们的系统还可以优化查询检索。盲索引无需解密即可对加密数据进行高效搜索，从而实现快速搜索，并为相关记录提供快速、直接的访问。对于非敏感数据检索，搜索查询会转换为 V 向量；这些向量使用词频-逆文档频率将文档转换为数值向量，同时隐藏其主要内容。我们使用 K 均值聚类来确定所需的 Q 所在的簇。CS 定位到该簇后，系统会比较搜索查询 Q 与该簇内文档的 V 向量之间的余弦相似度，并据此进行排序；与查询最相关的文档将优先返回。这种组合策略在安全消息传递、医疗记录或财务日志等场景中尤其有优势，因为在这些场景中，隐私和易用性都至关重要。

4. 结果

本节全面展示了我们提出的系统所取得的结果。我们的实验模型分为两个部分。第一部分涉及识别敏感数据（电子邮件、电话号码和信用卡号码）。第二部分则侧重于通过加密来保护敏感数据，并最终优化数据库。这种集成能够自动识别、加密和索引敏感字段，从而实现安全、高效且合规的查询。完成数据预处理步骤后（这一环节对取得更好的结果发挥了重要作用），我们使用深度学习算法、长短期记忆网络、卷积神经网络、深度置信网络、混合模型和基于 Transformer 的模型训练了所提出的模型。我们从准确率、精确率、召回率和 F1 分数等方面评估了这些模型的结果。我们提出的模型的第二阶段使用命名实体识别系统识别隐私感知数据，并采用 AES 加密对其进行加密。另一方面，为了优化数据库，我们对敏感信息使用盲索引，同时将非敏感文本向量化并聚类，以实现快速、高效的检索。此外，将命名实体识别与 NoSQL 数据库查询优化框架相结合，可确保高效查询敏感实体，同时维持较高的数据隐私水平。

4.1. 实验设置

测试和实验在一台独立计算机上进行，该计算机运行 Win- dows 11（64 位），配备 16 GB 内存。实现采用 Python 编程语言（版-

本 3.13.3）。Jupyter Notebook 这一种本地开发环境用于运行所有代码。使用 NumPy 和 Pandas 进行探索性数据分析。SpaCy 用于处理自然语言文本，包括命名实体、分词、词形还原、词向量化和词性标注。to_categorical 用于将文本转换为向量，而 pad_sequences 用于使所有输入长度相同。Scikit-learn 的 train_test_split 和 classification_report 用于拆分数据并生成分类报告。而 TensorFlow 这一用于分析神经网络的热门深度学习框架，则用于构建和测试我们的深度学习模型。matplotlib 和 seaborn 库用于生成图表和绘图。Hashlib 用于生成盲索引，并使用 Crypto 模块中的 AES 包。Cipher 包用于加密敏感信息。K 均值和 Scikit-learn 中的 TfidfVectorizer 用于聚类和提取文本特征。

4.2. 命名实体识别模型性能评估

我们采用了双向长短期记忆网络、卷积神经网络算法，以及结合长短期记忆网络-卷积神经网络、DBN-LSTM 和两种基于 Transformer 的算法 BERT 与 XLNET 的混合模型，以评估所提出系统的命名实体识别模型的性能。数据集按照 80:20 的比例进行划分，并采用 10 折交叉验证来验证数据集。在这些深度学习模型中，双向长短期记忆网络的参数信息包括：两个隐藏层，最大序列长度为 100，丢弃率在 0.2 至 0.5 之间（最优结果在 0.2 时获得），学习率为 0.001，预热比例为 0.1。批量大小在 16、32 和 64 之间变化；然而，批量大小为 32 时结果最佳，训练轮次范围为 25 至 40。嵌入维度值为 128，输入形状为（100,），并使用 Adam 优化器。另一方面，对于这两个混合模型，批量大小为 32，最大序列长度为 100，嵌入维度、训练轮次和输入形状与双向长短期记忆网络相同。不过，主要区别在于隐藏层。对于长短期记忆网络-卷积神经网络混合模型，第一个隐藏层是卷积层，滤波器大小为 64，卷积核数量为 3，并采用相同填充。长短期记忆网络是第二层，滤波器大小为 64；第三层是值为 128 的全连接层，激活函数为 ReLU。对于所提出模型的第二个混合模型，将长短期记忆网络与 DBN 结合，并使用 34 个循环单元；卷积层为第二层，滤波器大小为 64、卷积核大小为 3；第三层是大小为 128 的长短期记忆网络层。用于训练深度学习算法的超参数列表见表6。对于基于 Transformer 的模型，我们还使用了各种参数以获得最佳输出。BERT 和 XLNet 使用最大序列长度 128、批量大小 32 和学习率 2e-5 进行微调。两个模型均使用 Adam 优化器训练 30 个训练周期，并应用提前停止以防止过拟合。

表6：深度学习算法的超参数值

类别	双向长短期记忆模型	CNN+ LSTM 模型	DBM+ LSTM 模型
批量大小	16, 32, 64	32	32
Max Se-序列长度	100	100	100
第1层	双向长短期记忆网络 (64)	一维卷积 (64, kernel=3，相同)	全连接层（128）
第2层	时间分布式 全连接层（64，ReLU）	长短期记忆网络 （64）	一维卷积 (64, kernel=3)
第3层	N/A	全连接层 (128, ReLU)	长短期记忆网络 (128)
嵌入层 Dim	128	128	128
优化器	adam	adam	adam
训练轮次	30	25	40
标签 Format	填充，int，expand_维度(y, -1)	填充，整数，不扩展展_维度	填充、整数，不扩展展_维度
输入形状	(100,)	(100,)	(100,)

4.2.1. 双向长短期记忆网络的性能评估

表7展示了双向长短期记忆网络算法在前 25 个训练周期中的前 10 个训练周期的结果。从表中可以看出，训练准确率在经过 4 个训练周期后达到 85%，而在第一个训练周期中为 68%。训练准确率在第 3 个训练周期后稳步上升。另一方面，训练损失在第一个训练周期为 0.861，并在第二个训练周期显著下降至 0.396，随后在每次迭代中持续下降。经过 10 次迭代后，训练损失接近于零，而验证准确率在第一个训练周期从 59% 开始。尽管存在轻微波动，但其仍始终保持在 80% 以上。此外，验证损失从 2.1 开始，并随着模型在训练数据上的逐步改进，在第 10 个训练周期降至 0.9，而验证误差则缓慢上升。命名实体识别模型使用双向长短期记忆网络算法，在所有敏感实体上均表现出色（表8），包括信用_卡、统一资源定位符、电话、货币、社会安全号码以及非实体类别 O，其精确率、召回率和 F1 分数均超过 85%，表明模型能够完美识别这些词元。唯一的例外是电子邮件，其精确率为 83%、召回率为 77%、F1 分数为 82%（图6）。因此，结果表明该模型在识别敏感信息方面具有很高的精确性和稳健性，适用于包括隐私感知数据在内的任务

表7：BiLSTM算法在前10个训练周期的结果

轮次	训练准确率	训练损失	验证准确率	验证Loss
1	0.68	0.861	0.59	2.1
2	0.7345	0.396	0.61	1.6
3	0.7474	0.19	0.68	1.4
4	0.8511	0.17	0.72	1.3
5	0.8528	0.19	0.73	1.3
6	0.8562	0.16	0.75	1.25
7	0.8586	0.13	0.75	1.28
8	0.8589	0.13	0.8	1.18
9	0.8607	0.11	0.81	0.91
10	0.8908	0.08	0.8	0.9

表8：BiLSTM算法的实体性能指标

实体	精确率 (%)	召回率 (%)	F1分数（%）
B-信用_CARD	93	91	94
电子邮件起始标记	83	77	82
金额起始标记	96	88	90
电话号码起始标记	87	89	83
社会保障号码起始标记	89	86	79
网址起始标记	91	87	96
O	94	97	92

处理或安全信息提取。

4.2.2. 基于卷积神经网络-长短期记忆网络的性能评估

如表9所示，我们提出的卷积神经网络-长短期记忆网络混合模型在前10个训练周期展现出明显的学习趋势。第一个训练周期的训练准确率为61.15%，损失为2.9。经过10个训练周期后，准确率达到87.2%，损失则大幅降至1.62。另一方面，验证性能波动较大：第一个训练周期的准确率为77.05%，但在下一个训练周期骤降至69.5%，随后逐渐回升，同时损失升至1.8206，如图7所示。从表10可以看出，模型在不同实体上的表现各异；其中，CREDIT_CARD实体的结果最佳，其

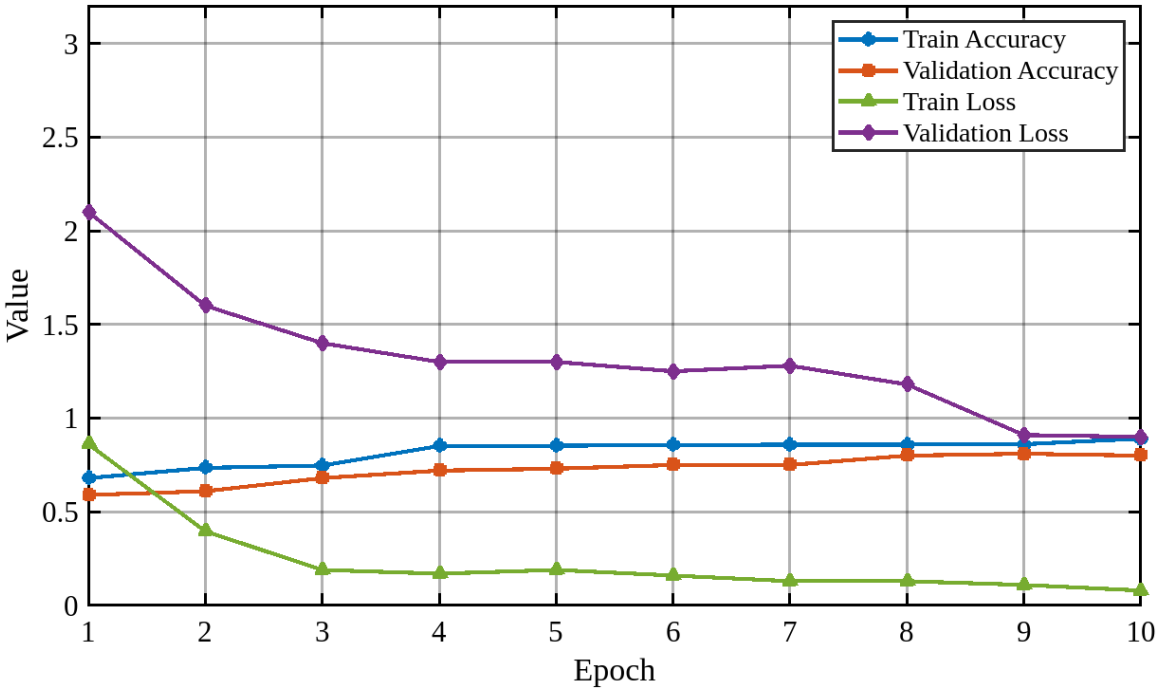


图6：使用 BiLSTM算法训练10个训练周期后四项指标的结果。

精确率为0.88，召回率为0.84，F1分数为0.87。此外，电话和货币类别的表现也不错，F1 分数分别为0.87和0.82。金额起始标记的召回率比精确率高0.1。电子邮件和统一资源定位符这两个实体的结果最差，F1 分数分别为0.72和0.67。非实体类别的F1 分数达到0.83，表现出色。

*4.2.3. DBN-LSTM*性能评估

如表11所示，在最初的30个训练周期中的前10个训练周期内，混合模型 DBN-LSTM 在几乎所有实体上的表现都十分出色，尤其是在学习和泛化能力方面。模型最初的训练准确率相对较低，为61.48%，对应的损失为2.2，之后持续上升（见图8）。到第10个训练周期时，模型的验证性能达到86.11%。值得注意的是，从第2个训练周期开始，验证损失显著下降，到最后一个训练周期时降至0.38。表12中不同实体的评估结果凸显了模型性能的显著差异。格式独特且结构明确的实体，例如 CREDIT_CARD 和 PHONE，取得了最高的性能，F1 分数分别为0.96和0.89。同样，MONEY 的 F1 分数为0.75，SSN 的 F1 分数为0.74，结果令人满意。另一方面，EMAIL 的 F1 分数为0.68，召回率相对较低，仅为0.64，这表明模型

表9：CNN-LSTM算法前10个训练周期的结果

训练轮数	训练准确率	验证准确率	训练损失	验证Loss
1	0.6015	0.7705	2.9	2.6
2	0.6328	0.695	2.9	2.53
3	0.6977	0.7133	2.7	2.512
4	0.721	0.7542	2.61	2.3
5	0.748	0.7562	2.65	2.28
6	0.824	0.7365	2.3	1.90
7	0.851	0.7072	2.1	1.8206
8	0.853	0.7348	1.8	1.3042
9	0.858	0.7884	1.7	1.1872
10	0.872	0.8439	1.62	1.1035

表10：CNN-LSTM算法的实体性能指标

实体	精确率	召回率	F1分数
信用起始标记_CARD	0.88	0.84	0.87
电子邮件起始标记	0.63	0.67	0.72
金额起始标记	0.86	0.88	0.87
电话号码起始标记	0.79	0.68	0.73
社会保障号码起始标记	0.60	0.68	0.67
网址起始标记	0.79	0.80	0.82
O	0.87	0.91	0.83

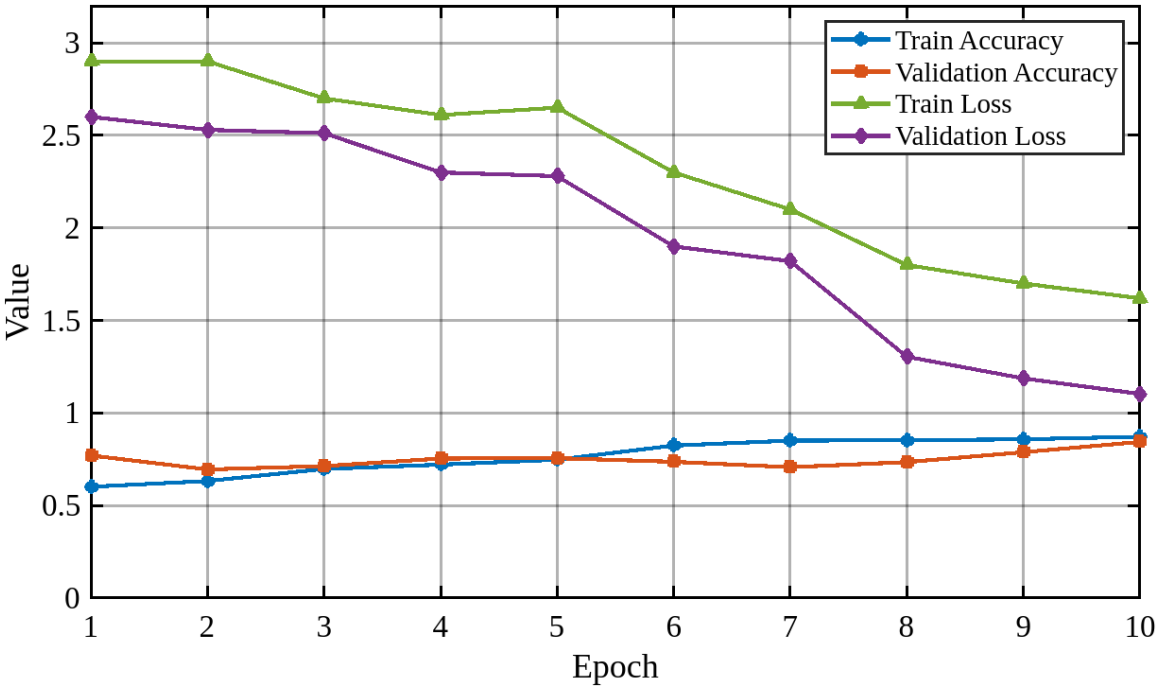


图7：使用 CNN-LSTM算法训练10个训练周期后四项指标的结果

难以识别所有相关的电子邮件实体。最后，代表非实体词元的 O标签取得了较高的 F1 分数（ 0.79 ），这表明其在区分实体文本和非实体文本方面表现稳定。

4.2.4. 所有算法的最终结果

要更改默认设置，请按如下方式调整模板。表13比较了若干深度学习算法和两种预训练Transformer模型（如 BERT 和 XLNET）完成训练后的最终结果。DBN与LSTM表现最佳，准确率（93%）、精确率（94%）、召回率（93%）和F1分数（93%）均为最高。LSTM模型也表现出色，尤其是在召回率和精确率方面（分别为93%和92%）。相比之下，卷积神经网络的准确率略高，为90%，但精确率和F1分数分别低了2%和10%；CNN-LSTM混合模型的表现较为均衡（准确率为86%，F1分数为82%），如图9所示。与其他深度学习模型相比，基于 Transformer 的模型（如 XLNet 和 BERT）准确率较低。XLNet在所有指标上都略胜于BERT，尤其是F1分数（83.4%对81%）。在表现较好的模型中，BERT的准确率最低，为79%。

表11：DBN-LSTM算法前10个训练周期的结果

轮次	训练 准确率-racy	训练损失	验证准确率-确率	验证 Loss
1	0.6148	2.2086	0.6319	1.8002
2	0.7066	1.5223	0.7778	0.8781
3	0.7785	1.3961	0.8264	0.6015
4	0.8171	1.1881	0.8611	0.5367
5	0.8670	0.9737	0.8611	0.4999
6	0.8566	0.92445	0.8819	0.4554
7	0.8927	0.9107	0.8750	0.4398
8	0.8802	0.882	0.8819	0.4058
9	0.8822	0.859	0.8958	0.3817
10	0.9019	0.8133	0.8611	0.3837

表12：DBN-LSTM算法的实体性能指标

实体	精确率	召回率	F1分数
B-信用 _CARD	0.95	0.98	0.96
电子邮件起始标记	0.72	0.64	0.68
金额起始标记	0.91	0.87	0.89
电话号码起始标记	0.84	0.66	0.74
社会保障号码起始标记	0.62	0.63	0.62
网址起始标记	0.73	0.77	0.75
O	0.76	0.82	0.79

表13：不同算法的性能比较

算法	准确率 (%)	精确率 (%)	召回率 (%)	F1分数 (%)
双向长短期记忆网络	89	92	93	91
卷积神经网络-长短期记忆网络	86	84	82	82
DBN-LSTM	93	94	93	93
XLNET	81	83	85	83.4
BERT	79	82	82	81
双向长短期记忆网络	89	92	93	91
卷积神经网络-长短期记忆网络	86	84	82	82

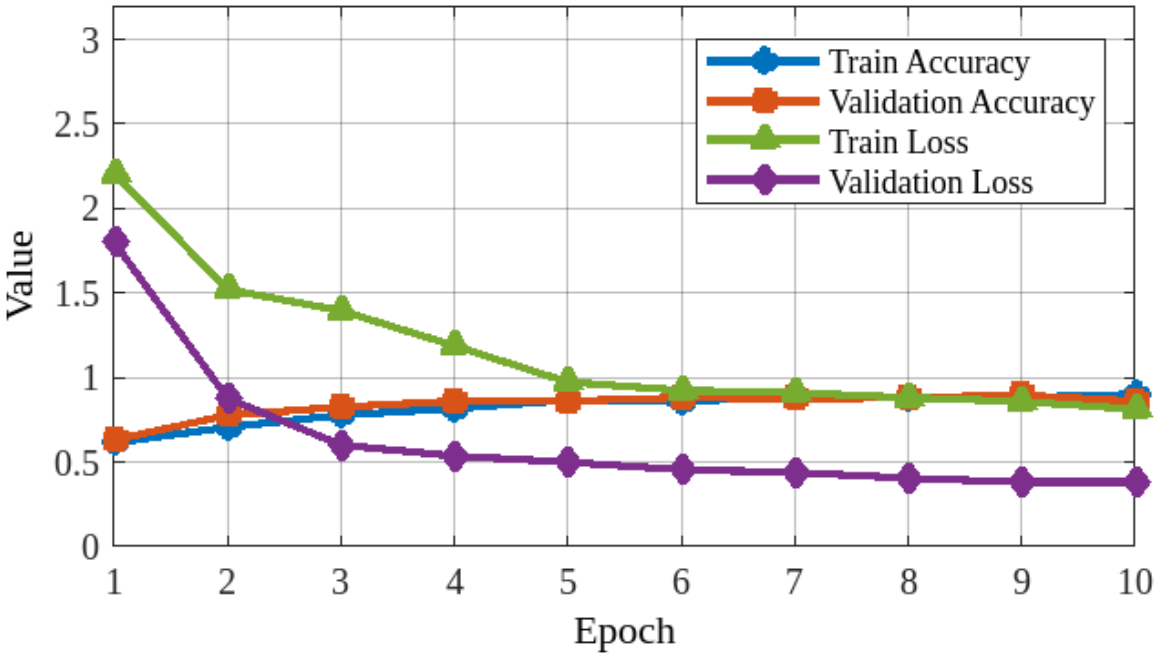


图8：使用DBN-LSTM算法训练10个训练周期后的四项指标结果

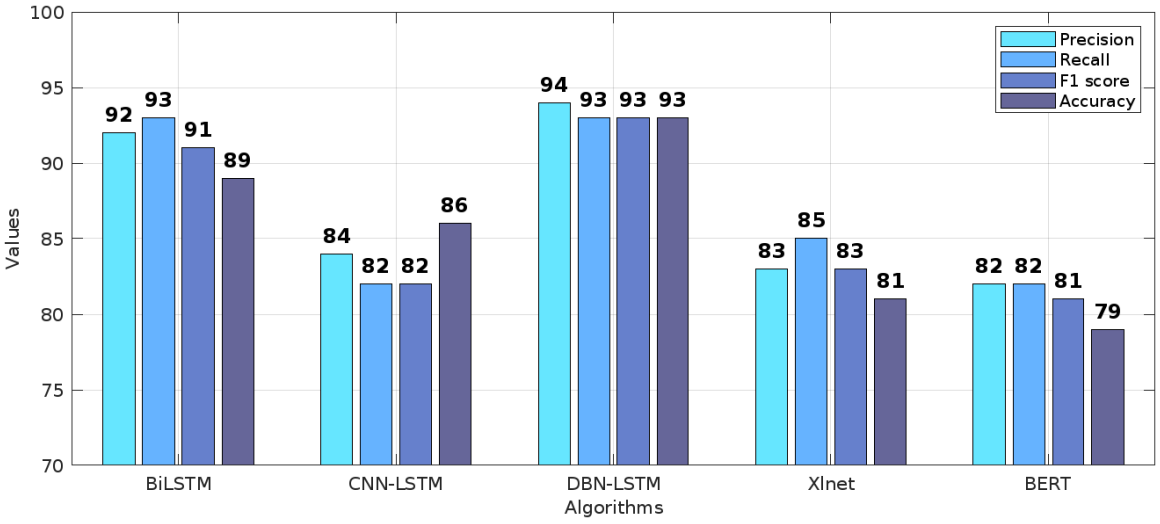


图9：命名实体识别模型检测模式中实体的五种模型性能比较

4.3. 数据库优化性能评估

所提出的系统显著提升了数据库优化性能：通过盲索引加速加密查询，并借助高效聚类和用于排序搜索的余弦相似度分数，提升非敏感数据的检索效率。此外，该系统还针对敏感信息和非敏感信息提出了高效的优化技术。

4.3.1. 敏感信息优化

采用 16 字节进行对称加密、使用 32 字节构建盲索引后，我们通过直接搜索和使用索引搜索敏感信息，验证了系统的有效性。

图10展示了加密时间（AES）、盲索引生成时间（HMAC）及其总成本如何随数据集大小变化。由于分组密码运算开销较大，AES 加密耗时更长，而 HMAC 哈希运算则更快。总构建成本呈线性增长，体现出可预测的可扩展性。

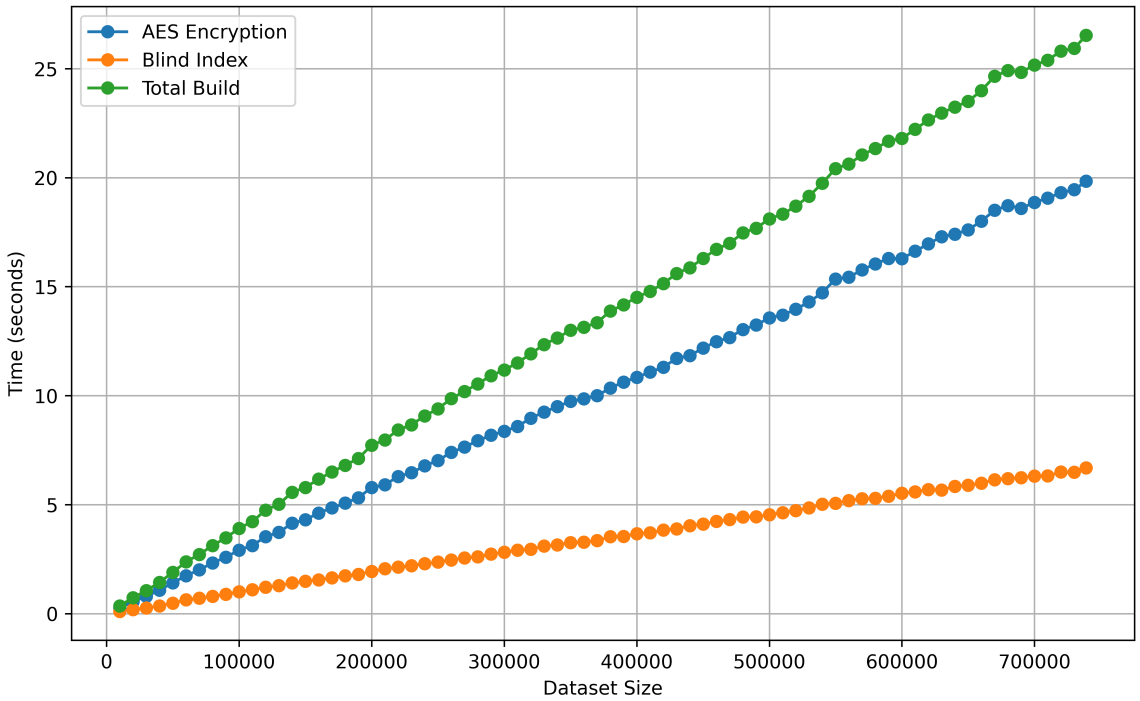


图10：构建时间与数据集大小的关系（AES、盲索引、总构建时间）

图 11 对比了两种搜索策略。直接查找曲线显示了在数据集中扫描明文数据所需的时间，这种方式采用线性搜索。随着数据集增大，搜索时间缓慢增加，体现了典型的基于列表的

查找行为。相反，盲查找曲线（使用存储在集合中的预计算 HMAC 值）几乎保持恒定，不受数据集大小影响。这是因为盲索引使用 Python 集合数据结构，其复杂度为 O(1)。两条曲线之间的差距凸显了盲索引的性能优势，尤其是在大规模数据集中，传统的明文扫描会变得越来越低效。

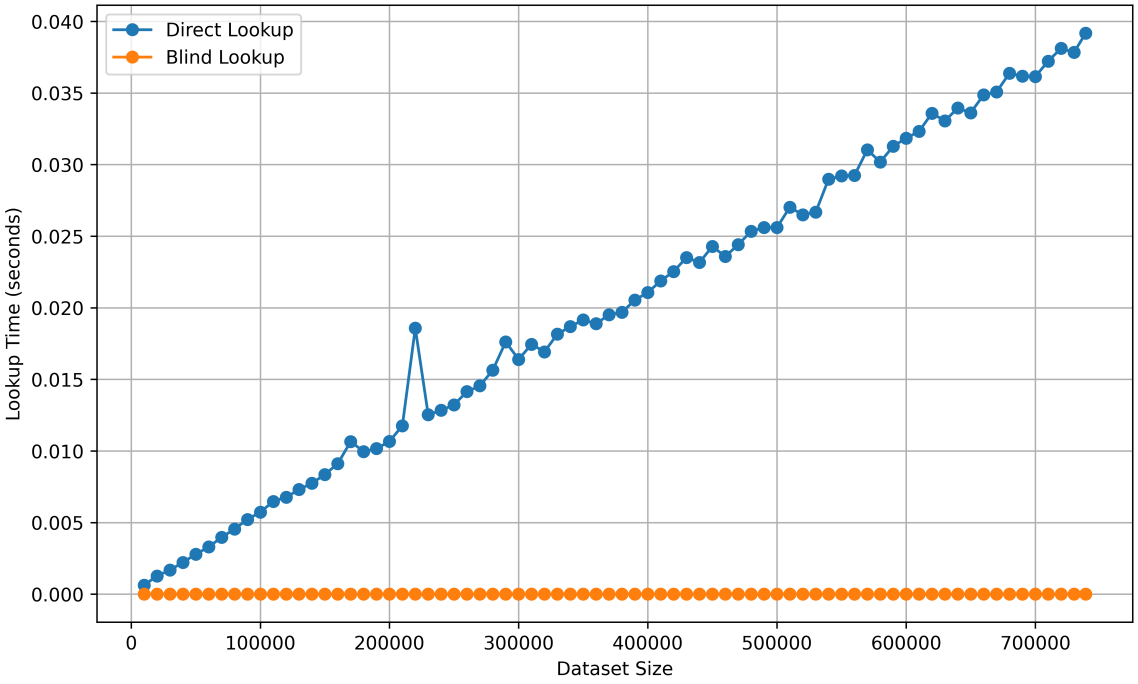


图 11：数据集的直接查找与盲查找耗时

4.3.2. 非敏感信息优化

图 12 比较了多种普通文本下聚类搜索与普通搜索的耗时。总体而言，普通搜索方法在大多数词条上的耗时更长，这表明它需要花费更多时间来检索向量形式的非敏感信息。GDPR 合规性查询在普通搜索中的得分为 0.60，而在聚类搜索中的得分为 0.27，耗时不到前者的一半。密码策略查询也呈现出显著差距（0.58 对 0.35）。对于每个搜索查询，聚类搜索的耗时都更短，而普通搜索则耗时显著更长。此外，对于“信用卡”“GDPR”和“用户同意”等查询字符串，簇中的余弦相似度分数分别展示于图 14、图 15 和图 13 中，最多显示 10 个匹配项；我们的系统会返回相似度分数最高的结果，从而实现快速响应并获得准确结果。

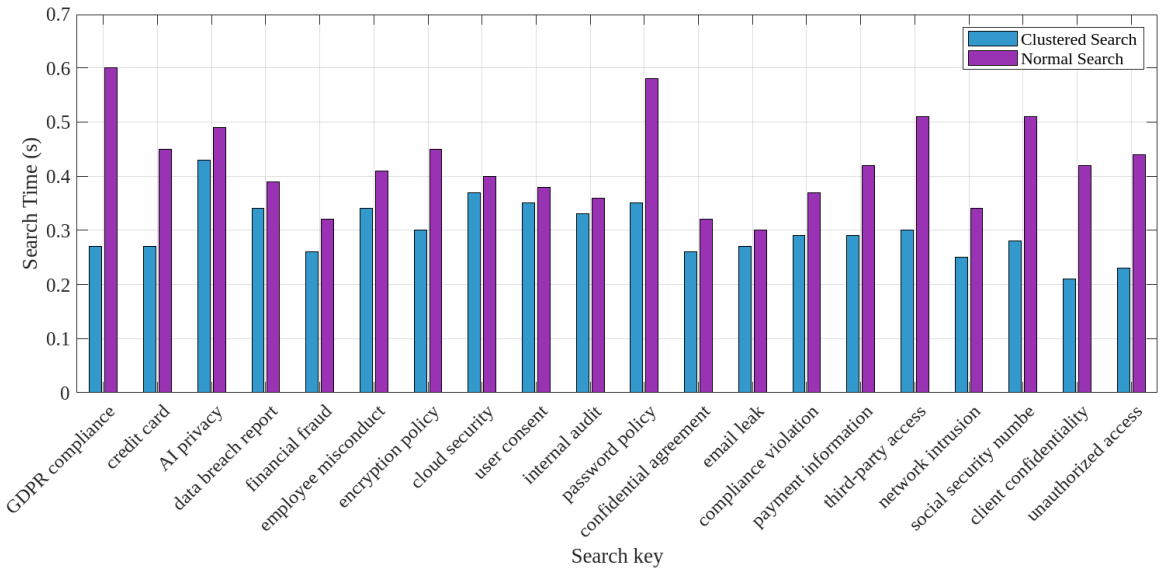


图 12：聚类搜索与普通搜索的非敏感信息比较

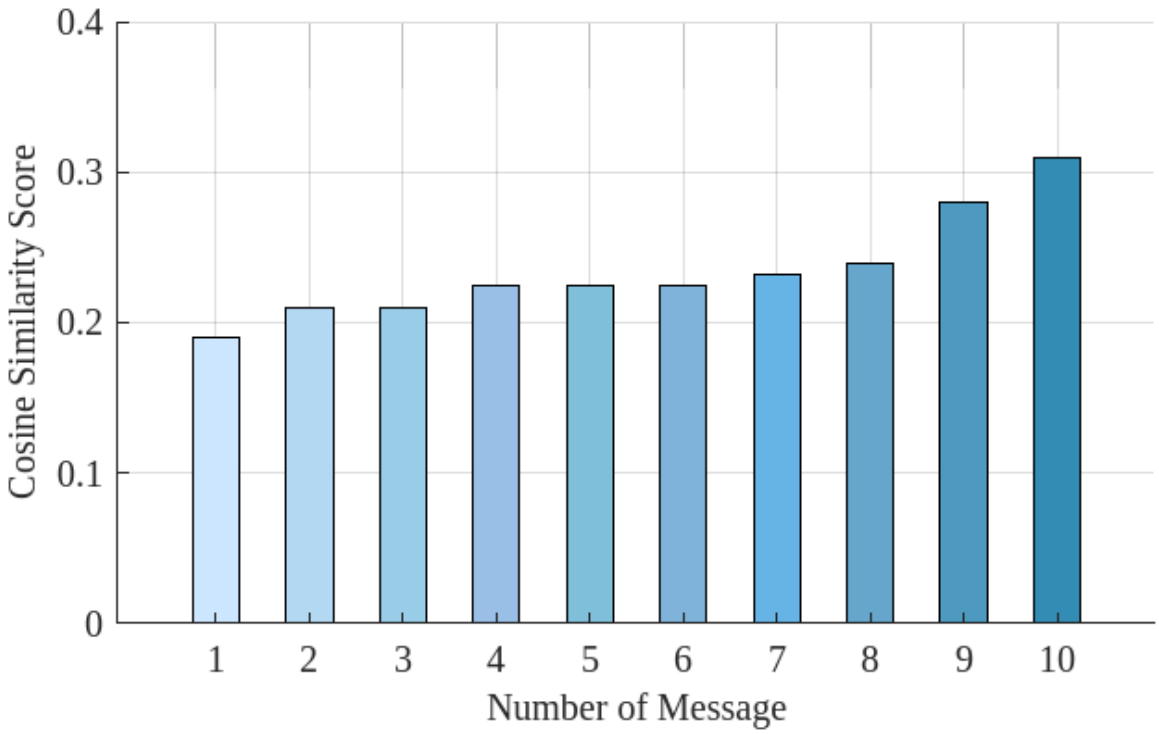


图 13：使用 DBN-LSTM算法训练10个训练周期后的四项指标结果

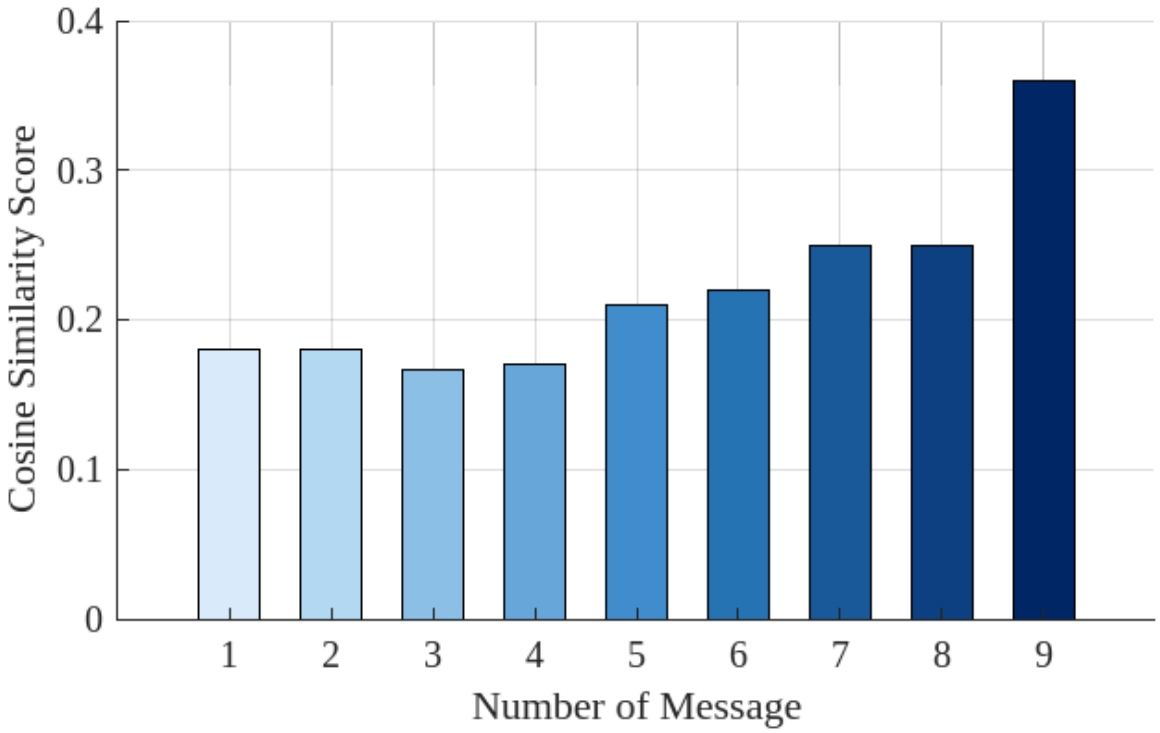


图 14：使用 DBN-LSTM算法训练 10个训练周期后得到的四项指标结果

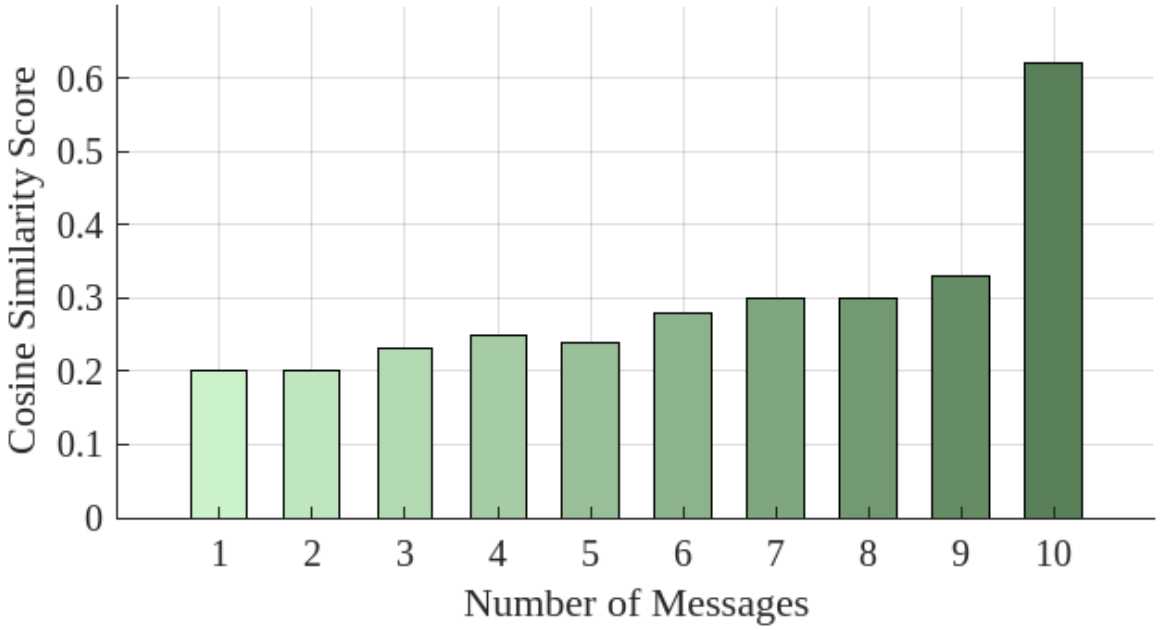


图 15：使用 DBN-LSTM算法训练 10个训练周期后得到的四项指标结果

5. 讨论

通过对多种模型在多项性能指标上的比较分析，可以深入了解其表现。双向长短期记忆模型表现出稳定的学习能力和出色的泛化能力。卷积神经网络-长短期记忆网络利用卷积层提取局部空间特征，并通过长短期记忆网络层提取时间特征。与仅使用卷积神经网络的模型相比，这种组合模型表现更佳；仅使用卷积神经网络的模型缺乏稳定性，这或许是因为难以平衡特征抽象与时间学习。相比之下，得分最高的模型是DBN-LSTM，其准确率为93%，精确率为94%，F1 分数为93%。这种多阶段预训练方法加深了学习并实现正则化，使模型能够学习更抽象、抗噪的特征，从而提升了其在所有指标上的表现。XLNET的表现令人满意。不过，其表现不佳可能是由于缺乏微调，或预训练任务与数据集的领域特定组织结构不匹配。BERT是另一种Transformer模型，在本研究中的准确率最低，F1分数也相对较低。尽管具备全局注意力机制，BERT仍因领域特定预训练不足且数据集规模相对较小而表现欠佳，因此未能有效泛化。从表14可以看出，我们提出的模型表现出色。我们提出的查询优化模型检索加密值所需的时间显著缩短，尤其是在检查电子邮件地址时。使用HMAC哈希在集合中查找电子邮件的时间复杂度为O(1)。这种差异体现了盲索引在可扩展性方面的巨大优势。在10万条记录中进行直接搜索需要5至20毫秒（耗时会随数据规模增加）；盲索引搜索仅需0.0001毫秒（耗时恒定，处于微秒范围）。因此，使用盲索引可使查找速度提升50,000至200,000倍。此外，我们还通过词频-逆文档频率向量化、聚类和搜索优化了非敏感信息。与标准搜索相比，采用这些机制可以更快地检索非敏感信息。为验证所提出的隐私感知NoSQL查询优化模型，出于多项实际和技术方面的考虑，我们选择了安然电子邮件数据集，而非真实的NoSQL数据集。首先，包含大量敏感数据的NoSQL数据库往往具有机密性、专有性，或因数据安全政策和隐私法规而无法访问。即使能够获取此类数据集，也需要支付高昂的会员费用、取得法律许可，或使用专门环境来运行高级NoSQL功能，这使得学术研究中的探索成本过高。相比之下，安然数据集 [81] 可免费获取，包含大量半结构化和非结构化数据（70万封电子邮件），并含有与NoSQL环境相关的真实敏感信息（例如姓名、电子邮件地址、日期）。此前的研究，例如 [43] [82][83], ，已成功利用安然语料库验证隐私感知与优化系统，进一步证明了该数据集可作为安全且优化的数据访问领域的标准基准。

表14：与最先进的方法进行比较

论文	准确率 (%)	精确率 (%)	召回率 (%)	F1分数 (%)
Wilkho等人 2025 [31]	N/A	93.60	91.74	92.66
Gao等人 [32] 2025	N/A	85.33	86.56	85.94
H. Cho等人 2019 [33]	N/A	96.05	95.55	96.29
Zhou等人 [34] 2025	89.09	89.09	89.49	89.18
Serere等人 [35] 2025	N/A	91	90	90
G.等人 [36] 2025	N/A	N/A	N/A	87.2
Wu 等人 2018 [37]	72	75.4	73.3	74.3
Zhang 等人 [38] 2025	N/A	97.36	75.22	84.87
Misra等人 [47]	89.4	94.7	89.4	91.1
我们提出的模型	93.2	94	93	93

通过利用基于深度学习的命名实体识别来识别敏感信息，结合加密和盲索引，并处理非敏感信息以实现快速检索，本研究提出了一种兼顾可扩展性与隐私保护的NoSQL查询优化方法，从而能够安全、高效地访问敏感数据。将DBN-LSTM等先进神经网络模型与保护隐私的加密搜索协议相结合，不仅提高了准确率，也加快了检索速度，使NoSQL系统更加安全、智能。所提出的隐私感知型NoSQL查询优化框架可为各类用户带来多层面的益处。该系统通过自动识别敏感信息并为开发人员优化相关处理，简化了安全数据集成。对于组织和服务提供商而言，该系统能够提升查询性能并降低数据泄露的风险。它还能改善用户的隐私保护和数据主权，确保其敏感数据和个人数据得到最高级别的安全保护，并以合乎伦理的方式处理。

6. 结论

本文提出了隐私保护仿真模型，通过识别敏感信息来优化数据库。其创新之处在于构建了一个系统：使用自然语言处理领域的命名实体识别模型，从NoSQL查询中自动识别敏感信息，即电子邮件、电话号码、信用卡号、余额和统一资源定位符；使用高级加密标准对其加密，并创建盲索引以实现高效搜索；生成向量化查询以保护非敏感数据；并对其进行聚类优化。在深度学习算法中，DBN-LSTM的表现最佳，准确率为93%，精确率为94%，召回率为93%，F1分数为93%。数据库优化则通过安然数据集进行验证。我们发现，所提出的系统执行时间更短，且比大多数竞争系统更安全。

然而，这些方法存在一些局限性。我们提出的命名实体识别模型只能识别数据集所定义的固定实体集合。文中未讨论涉及多个非敏感信息片段的复杂查询。另一个局限在于空间复杂度；存储加密数据、索引编号和向量值可能会占用存储空间。不过，存储空间略有增加是可以接受的，因为我们提出的系统在通过索引和聚类实现优化的同时，提供了双层安全保护。安全性和效率至关重要，尤其是在数据库中包含敏感信息时。

作为未来工作的一部分，我们将通过解决这些局限性来改进模型。此外，还将实现复杂但被忽视的聚合操作，例如聚合或连接。目前，所提出的方案仅表明存储和检索速度更快且更安全；未来将实现向量化数据检索以及转换回原始形式。此外，我们

将使用更多样化的聚类算法来改进聚类。网络以及企业和其他行业中可获取的敏感信息数量正在增加。我们提出的模型迫切需要在优化的同时保护敏感信息和个人数据。

参考文献

[1] H. Vera-Olivera、R. Guo、R. C. Huacarpuma、A. P. B. Da Silva、A. M. Mariano 和 M. Holanda，“数据建模与 NoSQL 数据库——系统映射综述”，*ACM Computing Surveys (CSUR)*，第 54 卷，第 6 期，第 1–26 页，2021年。[2] W. Khan、T. Kumar、C. Zhang、K. Raj、A. M. Roy 和 B. Luo，“SQL 与 NoSQL 数据库软件架构的性能分析与评估——系统文献综述”，*Big Data and Cognitive Computing*，第 7 卷，第 2 期，第 97 页，2023年。[3] R. Buyya、J. Broberg、R. Buyya、A. Goscinski、A. GoÃ、A. GoÃ、A. Goâsciânski 和 A. M. Goscinski，云计算。Wiley Online Library，2011年。[4] M. N. Sadiku、S. M. Musa 和 O. D. Momoh，“云计算：机遇与挑战”，*IEEE Potentials*，第 33 卷，第 1 期，第 34–36 页，2014年。[5] B. Alouffi、M. Hasnain、A. Alharbi、W. Alosaimi、H. Alyami 和 M. Ayaz，“云计算安全系统文献综述：威胁与缓解策略”，*IEEE Access*，第 9 卷，第 57 792–57 807 页，2021年。[6] F. K. Parast、C. Sindhav、S. Nikam、H. I. Yekta、K. B. Kent 和 S. Hakak，“云计算安全：基于服务的模型综述”，*Computers &Security*，第 114 卷，第 102580 页，2022年。[7] Y. S. Abdulsalam 和 M. Hedabou，“云计算中的安全与隐私：技术综述”，*Future Internet*，第 14 卷，第 1 期，第 11 页，2021年。[8] G. A. Pimenta Rodrigues、A. L. Marques Serrano、A. N. Lopes Espiñeira Lemos、E. D. Canedo、F. L. L. d. Mendonça、R. de Oliveira Albuquerque、A. L. San- doval Orozco 和 L. J. García Villalba，“从全局视角理解数据泄露：事件可视化与数据保护法综述”，*Data*，第 9 卷，第 2 期，第 27 页，2024年。[9]E. Bonnie，“110+ 最新数据泄露统计数据 [更新于 2025],”取自 https://secureframe.com/blog/data-breach-statistics，2025，访问日期：2025年1月3日。

[10] G. Viswanath 和 P. V. Krishna，“用于多云环境中大数据存储安全的混合加密框架”，*Evolutionary intelligence*，第 14 卷，第 2 期，第 691–698 页，2021年。[11] M. B. Qureshi、M. S. Qureshi、S. Tahir、A. Anwar、S. Hussain、M. Uddin 和 C. Chen，“卸载至云端的智能系统数据安全加密技术”，*Symmetry*，第 14 卷，第 4 期，第 695 页，2022年。[12] B. Seth、S. Dalal、V. Jaglan、D.-N. Le、S. Mohan 和 G. Srivastava，“整合加密技术以实现云端安全数据存储”，*Transactions onEmerging Telecommunications Technologies*，第 33 卷，第 4 期，第 e4108 页，2022年。[13] M. U. Sana、Z. Li、F. Javaid、H. B. Liaqat 和 M. U. Ali，“利用神经网络和加密技术增强云计算安全性”，*IEEE Access*，第 9 卷，第 145 785–145 799 页，2021年。[14] Z. Nasar、S. W. Jaffry 和 M. K. Malik，“命名实体识别与关系抽取：最新进展”，*ACM Computing Surveys (CSUR)*，第 54 卷，第 1 期，第 1–39 页，2021年。[15] B. Jehangir、S. Radhakrishnan 和 R. Agarwal，“命名实体识别综述——数据集、工具和方法体系”，*Natural Language ProcessingJournal*，第 3 卷，第 100017 页，2023年。[16] J. Kim、D. Kim、Y. Song、H. Shim、H. Sandberg 和 K. H. Johansson，“加密控制方法比较及基于学习带错误的同态加密的动态系统教程”，***Annual Reviews in Control***，第 54 卷，第 200–218 页，2022年。[17] B. Pulido-Gaytan、A. Tchernykh、J. M. Cortés-Mendoza、M. Babenko、G. Rad- chenko、A. Avetisyan 和 A. Y. Drozdov，“采用同态加密的隐私保护神经网络：挑战与机遇”，***Peer-to-Peer Networking and Applications***，第 14 卷，第 3 期，第 1666–1691 页，2021年。[18] A. Maté、J. Peral、J. Trujillo、C. Blanco、D. García-Saiz 和 E. Fernández-Medina，“通过模型驱动现代化提升 NoSQL 文档数据库的安全性”，*Knowledge and Information Systems*，第 63 卷，第 8 期，第 2209–2230 页，2021年。[19] A. A. Alflahi、M. A. Mohammed 和 A. Alsammani，“通过提升非关系型数据库系统（NoSQL）中的数据一致性来增强数据库访问性能”，载于 *International Conference on Computational Science and ItsApplications*。Springer，2024年，第 194–205 页。

[20] J. Kossmann、T. Papenbrock 和 F. Naumann，《用于查询优化的数据依赖：综述》，*The VLDB Journal*，第 31 卷第 1 期，第 1–22 页，2022年。[21] M. Jungmair、A. Kohn 和 J. Giceva，《设计用于查询优化与编译的开放框架》，*Proceedings of the VLDB Endowment*，第 15 卷第 11 期，第 2389–2401 页，2022年。[22] B. Zheng、X. Li、Z. Tian 和 L. Meng，《基于自适应双熵遗传算法的分布式数据库查询优化方法》，*IEEEAccess*，第 10 卷，第 4640–4648 页，2022年。[23] A. Sieminski、A. Kozierkiewicz、M. Nunez 和 Q. T. Ha，智能信息与数据库系统的现代方法。Springer，2018年，第 769 卷。[24] W. Deng 和 L. Zheng，《分类器准确率评估是否总是需要标签？》，发表于*IEEE/CVF*计算机视觉与模式识别会议，2021年，第 15 069–15 078 页。[25] M. Dias、J. Boné、J. C. Ferreira、R. Ribeiro 和 R. Maia，《用于葡萄牙语敏感数据发现的命名实体识别》，*Applied Sciences*，第 10 卷第 7 期，第 2303 页，2020年。[26] J. Sun、Y. Liu、J. Cui 和 H. He，《基于深度学习的自然灾害命名实体识别方法》，*Scientific reports*，第 12 卷第 1 期，第 4598 页，2022年。[27] N. Ahmed、A. K. Saha、M. A. Al Noman、J. R. Jim、M. Mridha 和 M. M. Kabir，《人机交互中的基于深度学习的自然语言处理：应用、进展与挑战》，*Natural Language Processing Jour-nal*，第 9 卷，第 100112 页，2024年。[28] R. Sallauka、U. Arioz、M. Rojc 和 I. Mlakar，《用于低资源语言症状提取的弱监督多语言医学命名实体识别》，*Applied Sci-ences*，第 15 卷第 10 期，第 5585 页，2025年。[29] H. Zhang、J. Wang、X. Wang、X. Lü、Z. Guan、Z. Cai 和 H. Zhang，《CPS-LSTM：电力系统的隐私敏感实体自适应识别模型》。*Energies(19961073)*，第 18 卷第 8 期，2025年。[30] S. Wang 和 L. Sun，《基于外部上下文检索与对抗训练的汽车故障文本中文命名实体识别》，*Entropy*，第 27 卷第 2 期，第 133 页，2025年。

[31] R. S. Wilkho 和 N. G. Gharaibeh，《Ff-ner：一种用于采集有关山洪及相关基础设施影响的基于网络信息的命名实体识别模型》，《国际灾害风险降低期刊》，第 105604 页，2025年。[32] K. Gao、J. Zhou、Y. Chi 和 Y. Wen，《Tourismner：一种基于实体边界联合预测的旅游命名实体识别方法》，*Intelligent Sys-tems with Applications*，第 25 卷，第 200475 页，2025年。[33] H. Cho 和 H. Lee，《利用带有上下文信息的深度神经网络进行生物医学命名实体识别》，*BMC bioinformatics*，第 20 卷第 1 期，第 735 页，2019年。[34] Z. Zhou、L. Wei 和 H. Luan，《利用深度学习从建筑施工物体撞击事故中提取关键信息以进行命名实体识别》，*Au-tomation in Construction*，第 173 卷，第 106106 页，2025年。[35] Y. Liu、K. Zhang、R. Tong、C. Cai、D. Chen 和 X. Wu，《一种用于嵌套命名实体识别的两阶段边界增强对比学习方法》，*Expert Systems with Applications*，第 271 卷，第 126707 页，2025年。[36] B. Kruthika、D. Gupta、S. Srivastava 等，《评估用于网络安全命名实体识别的各种微调大语言模型》，*Procedia Computer Science*，第 258 卷，第 1933–1943 页，2025年。[37] Y. Wu、M. Jiang、J. Xu、D. Zhi 和 H. Xu，《使用深度学习模型进行临床命名实体识别》，发表于*AMIA*年度研讨会论文集，第 2017 卷，2018年，第 1812 页。[38] Y. Zhang、J. Liu、X. Zhong 和 L. Wu，《Seclmner：一种利用大语言模型增强多源网络安全数据命名实体识别的框架》，*Expert Systems with Applications*，第 271 卷，第 126651 页，2025年。[39] H. Wang、W. Yang、W. Feng、L. Zeng 和 Z. Gu，《基于少样本学习的威胁情报命名实体识别技术》，*Array*，第 23 卷，第 100364 页，2024年。[40] D. Demirol、R. Das 和 D. Hanbay，《一种利用 BERT 模型分析网络威胁情报数据的网络威胁分析系统新方法》，*Symmetry*，第 17 卷第 4 期，第 587 页，2025年。[41] L. Wiese、T. Waage 和 M. Brenner，《Clouddbguard：一种用于 NoSQL宽列存储加密数据存储的框架》，《数据与知识工程》，第 126 卷，第 101732 页，2020年。

[42] L. Du, K. Li, Q. Liu, Z. Wu, 和 S. Zhang, “支持布尔查询的动态多客户端可搜索对称加密”, *Information Sciences*, 第506卷, 第234–257页, 2020年。[43] W. He, Y. Zhang, 和 Y. Li, “支持排序搜索的快速可搜索对称加密方案”, *Symmetry*, 第14卷, 第5期, 第1029页, 2022年。[44] B. L. Elizabeth 和 A. J. Prakash, “面向云数据的可验证 top-k 可搜索加密”, $S\bar{}adhan\bar{}a$, 第45卷, 第1期, 第9页, 2020年。[45] Y. Ping, W. Song, Z. Zhang, W. Wang, 和 B. Wang, “一种基于加密云数据概率陷门的多关键词可搜索加密方案”, *Information*, 第11卷, 第8期, 第394页, 2020年。[46] X. Liu, G. Yang, W. Susilo, J. Tonien, X. Liu, 和 J. Shen, “面向分布式系统的隐私保护多关键词可搜索加密”, *IEEE Transac-tions on Parallel and Distributed Systems*, 第32卷, 第3期, 第561–574页, 2020年。[47] K. Mishra, H. Pagare, 和 K. Sharma, “面向金融文档个人身份信息检测与匿名化的混合式基于规则的自然语言处理和机器学习方法”, *Scientific Reports*, 第15卷, 第1期, 第22729页, 2025年。[48] S. C. Fanni, M. Febi, G. Aghakhanyan, 和 E. Neri, “自然语言处理”, 载于人工智能导论。Springer, 2023年, 第87–99页。[49] D. Khurana, A. Koli, K. Khatter, 和 S. Singh, “自然语言处理：最新进展、当前趋势与挑战”, *Multimedia tools and applications*, 第82卷, 第3期, 第3713–3744页, 2023年。[50] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, 和 A. Chupryna, “自然语言处理应用中预处理算法的有效性”, 载于*2020 IEEE*信息通信问题：科学与技术国际会议（*PIC S&T*）。IEEE, 2020年, 第187–191页。[51] C. P. Chai, “文本预处理方法比较”, 自然语言工程, 第29卷, 第3期, 第509–553页, 2023年。[52] J.-D. Kim, T. Ohta, 和 J. Tsujii, “用于从文献中挖掘生物医学事件的语料库标注”, *BMC*生物信息学, 第9卷, 第1期, 第10页, 2008年。[53] N. Alshammari 和 S. Alanazi, “不同标注方案对命名实体识别的影响”, *Egyptian Informatics Journal*, 第22卷, 第3期, 第295–302页, 2021年。

[54] D. Baviskar, S. Ahirrao, 和 K. Kotecha，“多布局发票文档数据集（midd）：用于命名实体识别的数据集”，*Data*，第6卷，第7期，第78页，2021年。[55] S. Styawati 和 K. Mustofa，“用于电影观点数据分类的支持向量机-萤火虫算法”，*IJCCS*（印度尼西亚计算与控制论系统杂志），第13卷，第3期，第219–230页，2019年。[56] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, 和 H. Veisi，“基于改进预训练词嵌入的情感分析”，*Expert Systems with Applications*，第117卷，第139–147页，2019年。[57] U. Krzeszewska, A. Poniszewska-Marańda, 和 J. Ochelska-Mierzejewska，“分类情境下向量化方法的系统比较”，*AppliedSciences*，第12卷，第10期，第5119页，2022年。[58] S. Banani, S. Thiemjarus, K. Wongthavarawat, 和 N. Ounanong，“一种通过蓝牙低功耗信标实现安全数据传输的动态轻量级对称加密算法”，*Journal of Sensor and Actuator Networks*，第11卷，第1期，第2页，2021年。[59] B. Langenberg, H. Pham, 和 R. Steinwandt，“降低将高级加密标准实现为量子电路的成本”，*IEEE Transactions onQuantum Engineering*，第1卷，第1–12页，2020年。[60] L. Vialar, J. Ménétrey, V. Schiavoni, 和 P. Felber，“Blindextee：一种面向可信执行环境支持的端到端加密数据库管理系统的盲索引方法”，载于分布式系统稳定、安全与保障国际研讨会。Springer，2024年，第260–276页。[61] Y. Hu, W. Li, M. Ma, N. Cao, Y. Liu, Z. Qin, 和 J. Wang，“通过盲索引存储实现对加密云数据的复杂搜索”，载于*2017 IEEE*并行与分布式处理及应用国际研讨会暨*2017 IEEE*普适计算与通信国际会议（*ISPA/IUCC*）。IEEE，2017年，第1–8页。[62] D. V. Subramanian, J. Thangakumar, M. Preetha, S. Ganga, 和 S. Sajeev，“使用词频-逆文档频率、潜在狄利克雷分配和加权关联矩阵进行文档相似度计算与排序”，载于*2024*年数据工程与智能计算系统进展国际会议（*ADICS*）。IEEE，2024年，第01–07页。[63] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, 和 W. Khan，“ 基于朴素贝叶斯方法，使用词频 - 逆文档频率和计数向量化器对电影评论进行情感分析”，*Social network analysis and mining*，第14卷，第1期，第87页，2024年。

[64] J. A. Hartigan 和 M. A. Wong，“算法 AS 136：一种 k 均值聚类算法”，《皇家统计学会杂志：*C*辑（应用统计）》，第 28 卷，第 1 期，第 100–108 页，1979 年。[65] A. Likas、N. Vlassis 和 J. J. Verbeek，“全局 k 均值聚类算法”，《模式识别》，第 36 卷，第 2 期，第 451–461 页，2003 年。[66] S. Na、L. Xumin 和 G. Yong，“k 均值聚类算法研究：一种改进的 k 均值聚类算法”，载于 *2010*年第三届智能信息技术与安全信息学国际研讨会。IEEE，2010 年，第 63–67 页。[67]K. P. Sinaga 和 M.-S. Yang，“无监督 k 均值聚类算法”，*IEEE Access*，第 8 卷，第 80 716–80 727 页，2020 年。[68] T. Thireou 和 M. Reczko，“用于预测真核生物蛋白质亚细胞定位的双向长短期记忆网络”，《*IEEE/ACM*计算生物学与生物信息学汇刊》，第 4 卷，第 3 期，第 441–446 页，2007 年。[69] P. Zhou、W. Shi、J. Tian、Z. Qi、B. Li、H. Hao 和 B. Xu，“用于关系分类的基于注意力的双向长短期记忆网络”，载于计算语言学协会第 *54* 届年会论文集（第 *2* 卷：短论文），2016 年，第 207–212 页。[70] K. O’Shea 和 R. Nash，“卷积神经网络简介”，*arXiv* 预印本 *arXiv:1511.08458*，2015 年。[71] Z. Li、F. Liu、W. Yang、S. Peng 和 J. Zhou，“卷积神经网络综述：分析、应用与展望”，《*IEEE*神经网络与学习系统汇刊》，第 33 卷，第 12 期，第 6999–7019 页，2021 年。[72] R. Yamashita、M. Nishio、R. K. G. Do 和 K. Togashi，“卷积神经网络：概述及其在放射学中的应用”，《影像洞察》，第 9 卷，第 4 期，第 611–629 页，2018 年。[73] Q. Zhang、M. Zhang、T. Chen、Z. Sun、Y. Ma 和 B. Yu，“卷积神经网络加速的最新进展”，《神经计算》，第 323 卷，第 37–51 页，2019 年。

[74] G. E. Hinton，“Deep 信念网络，”学者 *pedia*，第4卷，第5期，p. 5947，2009 年。

[75] A.-r. Mohamed、G. E. Dahl 和 G. Hinton，“使用深度信念网络进行声学建模”，《*IEEE*音频、语音与语言处理汇刊》，第20卷，第1期，第14–22页，2011年。[76] Y. Ji、Z. Zhou、H. Liu 和 R. V. Davuluri，“Dnabert：用于基因组DNA语言的预训练来自变换器的双向编码器表示模型”，《生物信息学》，第37卷，第15期，第2112–2120页，2021年。[77] F. Sun、J. Liu、J. Wu、C. Pei、X. Lin、W. Ou 和 P. Jiang，“Bert4rec：利用来自变换器的双向编码器表示进行序列推荐”，载于第*28*届*ACM*信息与知识管理国际会议论文集，2019，第1441–1450页。[78] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“Bert：用于语言理解的深度双向变换器预训练”，载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇和短篇论文），2019，第4171–4186页。[79]A. F. Adoma、N.-M. Henry 和 W. Chen，“BERT、RoBERTa、DistilBERT 和 XLNET 在基于文本的情绪识别中的比较分析”，载于*2020*年第*17*届国际计算机会议：小波、主动媒体技术与信息处理（*ICCWAMTIP*）。IEEE，2020，第117–121页。[80] P. Rajapaksha、R. Farahbakhsh 和 N. Crespi，“BERT、XLNET 还是 RoBERTa：检测标题党的最佳迁移学习模型”，*IEEE Access*，第9卷，第154 704–154 716页，2021年。[81] B. Klimt 和 Y. Yang，“安然语料库：用于电子邮件分类研究的新数据集”，载于欧洲机器学习会议。Springer，2004，第217–226页。[82] D. Cash、S. Jarecki、C. Jutla、H. Krawczyk、M.-C. Roşu 和 M. Steiner，“支持布尔查询的高可扩展性可搜索对称加密”，载于年度密码学会议。Springer，2013，第353–373页。[83] W. Zhang、Y. Lin、S. Xiao、J. Wu 和 S. Zhou，“云计算中面向多个数据所有者的隐私保护排序多关键词搜索”，《*IEEE*计算机汇刊》，第65卷，第5期，第1566–1577页，2015年。