University of Brasília

Institute of Exact Sciences
Department of Computer Science

Privacy-Preserving Techniques for preparing texts for use in artificial intelligence models: Leveraging Semantic Similarity with Vector Data Search and AI Agents

Daniel Linhares Lim-Apo

Dissertation submitted in partial fullfilment of conclusão do
Professional Master's Degree in Applied Computing

Advisor
Prof.a Dr.a Edna Dias Canedo

Brasília
2026

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Lim-Apo, Daniel Linhares
LL732pp Privacy-Preserving Techniques for preparing texts for use in artificial intelligence models: Leveraging Semantic Similarity with Vector Data Search and AI Agents / Daniel Linhares Lim-Apo; orientador Edna Dias Canedo. Brasília, 2026.
101 p.

Dissertação(Mestrado Profissional em Computação Aplicada) Universidade de Brasília, 2026.

1. Privacy Preserving Techniques. 2. Semantic Text similarity. 3. Differential Privacy and Rare Events. 4. Latent Dirichlet Allocation (LDA) Topic Modeling. 5. AI Agents and AI Training Data. I. Canedo, Edna Dias, orient. II. Título.

University of Brasília

Institute of Exact Sciences
Department of Computer Science

Privacy-Preserving Techniques for preparing texts for use in artificial intelligence models: Leveraging Semantic Similarity with Vector Data Search and AI Agents

Daniel Linhares Lim-Apo

Dissertation submitted in partial fullfilment of conclusão of
Professional Master's Degree in Applied Computing

Prof.a Dr.a Edna Dias Canedo (Advisor)
CIC/UnB

Prof. Dr. Laerte Peotta de Melo
PPCA/UnB

Dr. Davi Viana
Universidade Federal do Maranhão (UFMA)

Prof.a Dr.a Edna Dias Canedo
Coordinator of the Postgraduate Program in Applied Computing

Brasília, abril 23, 2026

Dedicatória

Dedico esta obra à minha mãe, Conceição, que me deu a vida e a sabedoria; ao meu falecido pai, Rudolf, cuja inteligência, paciência e resiliência me inspiraram; à minha amada esposa, Valéria, pelo seu amor infinito; e aos meus queridos e amados filhos, Yohan e Pedro Jorge; ao meu irmão André e sua família por seu companheirismo e apoio.

Dedico também esta obra àqueles que estão ao lado das máquinas, àqueles que se opõem a elas e àqueles que se tornaram um com elas.

I dedicate this work to my mother, Conceição, who gave me life and wisdom; my late father, Rudolf, whose intelligence, patience and resilience inspired me; my beautiful wife, Valéria, for her boundless love; and my wonderful and beloved sons, Yohan and Pedro Jorge; to my brother André and his family for their companionship and support.

I also dedicate this work to those who stand with the machines, those who stand against them, and those who have become one with them.

Agradecimentos

Meus sinceros agradecimentos a todos os meus amigos, colegas, professores e alunos, com quem aprendi cada vez mais sobre a vida, suas coisas e seus pensamentos. À minha querida esposa, Valéria, e aos meus maravilhosos filhos, Yohan e Pedro Jorge, com amor e gratidão. Um agradecimento especial e sincero à minha mãe, Maria Conceição Linhares, a melhor mãe do mundo, que se dedicou tanto à nossa criação. Ao meu falecido pai, Rudolf Egbertus Lim-Apo, tão paciente e inteligente, e a toda a minha família.

Um agradecimento especial à minha orientadora, Dra. Edna Canedo, por seu trabalho excepcional em me inspirar, capacitar e guiar verdadeiramente na produção de pesquisas que vão além.

E um agradecimento também a todas as pessoas que foram professores, colegas, alunos e amigos ao longo da minha vida.

My heartfelt thanks to all my friends, colleagues, professors, and students, from whom I have continuously learned more and more about life and its things and thoughts. To my dear wife, Valéria, and my wonderful children, Yohan and Pedro Jorge, with love and gratitude. With a special and heartfelt thank you to my mother, Maria Conceição Linhares, the best mother in the world, who dedicated so much of herself to raising us. To my late father, Rudolf Egbertus Lim-Apo, so patient and intelligent, and to all my family.

A special thank you to my advisor, Dr. Edna Canedo, for her outstanding work in inspiring, enabling, and truly guiding me toward producing research that goes beyond.

And a thank you also to all the people who have been teachers, colleagues, students, and friends throughout my life.

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), through Access to the Periodicals Portal.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Contexto: Processos que visam extrair valor da informação a partir de dados amazenados estão ganhando destaque. Entre os diversos tipos de dados não estruturados, os dados textuais constituem uma parcela significativa da informação produzida em contextos do mundo real. Considerações éticas e leis de proteção de dados aumentaram a pressão sobre a privacidade de conteúdo sensível. Os riscos de divulgação associados a dados textuais, considerando a privacidade diferencial, são influenciados pela raridade e similaridade dos textos dentro de um conjunto de dados. Textos raros podem aumentar a probabilidade de reidentificação. A Inteligência Artificial (IA) e o Aprendizado de Máquina (ML) têm demanda crescente por dados e, juntamente com a estatística e as técnicas clássicas de processamento de linguagem natural, essas técnicas estão sendo cada vez mais exploradas para implementar mecanismos de preservação da privacidade, oferecendo soluções técnicas para mitigar os riscos à privacidade. **Objetivo:** O objetivo é descobrir técnicas de ponta para a preservação da privacidade no processamento de dados textuais, permitindo o emprego de técnicas para proteger a privacidade em dados não estruturados, especificamente em textos e na implementação de técnicas de similaridade textual. **Método:** Para atingir esse objetivo, foi pesquisado em busca do estado da arte quanto a técnicas de preservação de privacidade, por meio de uma revisão bibliográfica, e o estudo propôs a aplicação de técnicas selecionadas. Os conceitos de privacidade diferencial, bancos de dados vetoriais, similaridade textual e eventos raros foram considerados na metodologia e nos estudos de casos propostos, juntamente com o uso de sistemas de IA multiagentes e LLMs. **Resultados:** Uma contribuição fundamental deste estudo foi identificar as técnicas de ponta para preservação da privacidade aplicadas na análise de dados textuais e similaridade de textos, incluindo como a Ciência de Dados, Modelos de Liguagem em Larga Escala(LLM) e a Inteligência Artificial (IA) baseada em agentes são utilizadas para implementar mecanismos de preservação da privacidade, bem como as técnicas empregadas para similaridade semântica e detecção de eventos raros em domínios textuais. Além disso, foram apresentadas aplicações práticas em dois estudos de casos para o uso desse conhecimento. **Conclusão:** Este estudo oferece uma síntese estruturada dos estudos existentes por meio de um Mapeamento Sistemático de Estudos (MSE) e uma

perspectiva prática através de estudos de caso, destacando técnicas de preservação da privacidade na análise de texto. Ele ressalta a possibilidade de usar métodos de similaridade semântica e representações vetoriais na identificação de eventos raros em contextos sob restrições de privacidade. A integração de Modelos de Linguagem (LLMs) e Agentes de Inteligência Artificial (IA) revela-se promissora, mas, por outro lado, apresenta desafios e complexidades específicos para o processamento com foco na privacidade, particularmente em áreas como segurança pública. Este estudo forneceu uma visão geral da implementação e do uso prático, aplicado em um estudo de caso, de técnicas de similaridade semântica entre textos, que, conforme revelado no MSE, possuem uma presença forte e consolidada na literatura. Dada a escassez de abordagens similares na literatura pesquisada, este trabalho ajuda a preencher esta lacuna e busca contribuir para pesquisas futuras focadas em conciliar métodos de IA com aplicações éticas e que preservem a privacidade.

Palavras-chave: Privacidade, Privacidade Diferencial, Similaridade Semântica de Textos, Eventos Raros, Agentes AI, LLM, Latent Dirichlet Allocation, LDA, Treinamento de modelos de AI

Resumo Expandido

Técnicas de Preservação de Privacidade para a preparação de textos para uso em modelos de inteligência artificial: Fazendo uso de Similaridade Semântica de Textos com Pesquisa de Dados Vetoriais e Agentes de IA

Introdução

Este trabalho apresenta uma visão geral sobre técnicas de preservação de privacidade aplicadas a dados estruturados e não estruturados, com foco especial em dados textuais. O estudo examina técnicas de privacidade diferencial, os riscos de reidentificação e o uso de modelos de Processamento de Linguagem Natural (PLN), Inteligência Artificial (IA), agentes de Inteligência Artificial (IA), modelos de linguagem de grande escala (LLMs) em tarefas de anonimização e detecção de eventos raros.

A pesquisa discute como os dados armazenados estão se tornando matéria-prima valiosa para modelos de machine learning, e como isso gera tensões entre utilidade dos dados e proteção da privacidade. O uso de dados sensíveis em IA acontece sujeito a exigências técnicas, éticas e legais, como robustez, transparência e conformidade regulatória com legislações como a General Data Protection Regulation (GDPR) (Europa) e a Lei Geral de Proteção de Dados Pessoais (LGPD) (Brasil).

Os principais desafios trabalhados foram: A privacidade em textos não estruturados, que podem conter informações sensíveis de forma implícita; A identificação de técnicas de preservação da privacidade em textos, que exige técnicas sofisticadas, além das tradicionais utlizadas em dados estruturados; A detecção de eventos raros, onde textos com características incomuns podem facilitar reidentificação; O trade-off entre privacidade e utilidade, pois maior proteção geralmente implica menor precisão nos modelos.

O estudo adota como metodologia as primeiras fases do processo Cross Industry Standard Process for Data Mining (CRISP-DM)[1, 2], ou seja, correspondendo às fases de Compreensão do Negócio, Compreensão dos Dados e Preparação dos Dados, com um estudo de caso que analisa a aplicação de técnicas de privacidade em dados textuais e estruturados.

Dado o escopo amplo dos dados não estruturados, o trabalho delimita seu foco especificamente para textos, discutindo os riscos de inferência de informações pessoais por

modelos de PLN mesmo em corpora aparentemente anonimizados. Aponta também que, em cenários como o de eventos raros, técnicas como representações vetoriais (embeddings) podem ajudar a medir o quão incomum um texto é, auxiliando na detecção de possíveis riscos à privacidade.

Por fim, destaca-se que, embora a anonimização seja atualmente uma abordagem bem adotada para mitigar riscos em aprendizado de máquina, técnicas como pseudoanonimização podem ser preferidas em contextos onde a reidentificação controlada é necessária.

Objetivo

O objetivo é descobrir técnicas de ponta, o estado da arte, para a preservação da privacidade no processamento de dados textuais, permitindo o emprego de técnicas para proteger a privacidade em dados não estruturados, especificamente em textos e na implementação de técnicas de similaridade textual.

Metodologia

Esta pesquisa foi conduzida de forma estruturada, dividida em duas fases distintas, com o objetivo de alcançar uma abordagem sistemática para a aplicação de técnicas de preservação da privacidade voltadas à anonimização de dados.

Fase 1 – Revisão da Literatura: A primeira etapa consistiu em um Mapeamento Sistemático de Estudos (MSE), abordando técnicas de preservação da privacidade, modelos de privacidade diferencial e metodologias de detecção de eventos raros. Também foram analisados riscos de reidentificação associados a quase-identificadores e a eficácia de modelos de inteligência artificial na proteção da privacidade. Além disso, investigou-se a aplicação de modelos de linguagem de grande porte (LLMs) na anonimização de dados. Esta fase resultou em uma síntese do conhecimento existente, identificando lacunas e melhores práticas para orientar a seleção de técnicas de privacidade na fase seguinte.

Fase 2 – Estudos de Caso: Com base na revisão anterior, foram selecionadas técnicas e conceitos adequados para viabilizar a aplicação prática em estudos de caso.

No primeiro estudo de caso, foram explorados métodos para detecção de eventos raros em dados textuais por meio de análise vetorial, PLN e agentes de IA. Com uma abordagem integrando redução de dimensionalidade, banco de dados vetorial e métricas de similaridade, como distância de cosseno para detectar outliers semânticos e textos semelhantes.

No segundo estudo de caso, foi explorado o uso da Alocação Latente de Dirichlet (LDA) em uma proposta prática na perspectiva da Engenharia de Requisitos (ER) na análise de dados de treinamento de IA.

Resultados e Discussão

Uma das principais contribuições deste estudo consistiu na identificação e sistematização de técnicas contemporâneas voltadas à preservação da privacidade no contexto da

análise de dados textuais e da avaliação de similaridade entre textos. O trabalho examinou como abordagens baseadas em Ciência de Dados, Modelos de Linguagem de Grande Escala (LLMs) e Inteligência Artificial (IA) orientada por agentes têm sido empregadas na implementação de mecanismos de proteção da privacidade. Foram também analisadas as estratégias utilizadas para mensuração de similaridade semântica e identificação de eventos raros em domínios textuais.

Complementarmente, foi apresentado um primeiro estudo de caso ilustrando a aplicação prática desses conhecimentos em um cenário específico. Esta seção do estudo examinou o papel duplo dos sistemas de detecção de eventos raros, observando que, embora esses métodos aprimorem a capacidade de identificar padrões textuais únicos ou anômalos, eles também levantam preocupações relacionadas à privacidade, especialmente no que diz respeito ao potencial de reidentificação de pessoas envolvidas no caso de os fatos da vida serem eventos raros.

Com base nas técnicas identificadas no MSE, o segundo estudo de caso investigou a aplicação da Alocação Latente de Dirichlet (LDA) em uma proposta prática, sob a perspectiva da Engenharia de Requisitos (ER), voltada à análise de dados de treinamento de sistemas de IA.

A pesquisa contribui para as discussões em andamento nas áreas de processamento de linguagem natural, análise de dados com foco na privacidade e similaridade semântica baseada em vetores. Ela se alinha a esforços já presentes na literatura, citadas no MSE ao abordar as implicações associadas à raridade textual em contextos de dados sensíveis.

Conclusões

Este estudo apresenta uma síntese organizada da produção científica existente por meio de um MSE, complementado por uma abordagem prática baseada em estudos de caso, com ênfase nas técnicas de preservação da privacidade aplicadas à análise de textos. A investigação aborda o uso de métodos de similaridade semântica e representações vetoriais como ferramentas para a identificação de eventos raros em cenários caracterizados por restrições de privacidade.

A incorporação de Modelos de Linguagem de Grande Escala (LLMs) e de Agentes baseados em IA é discutida como uma abordagem promissora, embora acompanhada de desafios e complexidades particulares quando aplicada a contextos que demandam tratamento sensível dos dados, como na área de segurança pública. O estudo também oferece uma visão geral da aplicação prática de técnicas de similaridade semântica entre textos, cuja presença na literatura científica tem sido cotidianamente bem documentada, conforme evidenciado no MSE conduzido.

Considerando a escassez de trabalhos que integram essas abordagens sob a perspectiva da privacidade, este estudo busca contribuir para a consolidação de uma linha de pesquisa

voltada à integração de métodos baseados em IA com práticas éticas e orientadas à proteção de dados sensíveis.

Palavras-chave: Privacidade, Privacidade Diferencial, Similaridade Semântica de Textos, Eventos Raros, Agentes AI, LLM, Latent Dirichlet Allocation, LDA, Treinamento de modelos de AI

Abstract

Context: Processes that aim to extract value from stored data are gaining prominence. Among various types of unstructured data, textual data constitutes a significant proportion of the information produced in real-world settings. Ethical considerations and data protection laws have increased the pressure over the privacy of sensitive content. The disclosure risks associated with textual data, considering differential privacy, are influenced by the rarity and the similarity of texts within a dataset. Rare texts can increase the likelihood of re-identification. Artificial Intelligence (AI) and Machine Learning (ML) have growing demand for data and side by side with statics and classic natural language processing techniques, those techniques are increasingly being explored for implementing privacy-preserving mechanisms, offering technical solutions to mitigate privacy risks. **Goal:** The objective was to identify state-of-the-art techniques for privacy-preserving processing of textual data. The focus is on enabling the application of methods that protect privacy in unstructured data, particularly text, and in the implementation of text similarity approaches. **Method:** To achieve this goal, state-of-the-art privacy preservation techniques were researched, in a literature review, and the study proposed the application of selected techniques. The concepts of differential privacy, vector databases, text similarity and rare events were taken into account in the proposed methodology and case study, along with the use of multi-agent Artificial Intelligence (AI) systems and Large Language Models (LLMs). **Results:** A key contribution of this study was to identify the state of art techniques for the privacy-preserving that are applied in textual data analysis and text similarity, including as how Data Science, LLM and Agent-Based AI techniques are used to implement privacy-preserving mechanisms and the techniques that are employed for semantic similarity and rare events detection in text domains. And also, a purposed applications in two case studies for the use of that knowledge. **Conclusion:** This study offers both a structured synthesis of existing studies through a Systematic Mapping Study (SMS) and a practical perspective via a case study, highlighting privacy-preserving techniques in text analysis. It highlights the possibility of using semantic similarity methods and vector-based representations in identifying rare events in contexts under privacy constraints. The integration of LLMs and AI agents reveals promising but in other hand

there are specific challenges and complexity for privacy-aware processing, particularly in areas like public security. This study provided an overview of the implementation and practical use, applied in a case study, of semantic similarity techniques between texts, which were revealed in Systematic Mapping Study (SMS) to have a strong and mature presence in the literature and a second case study explored the use of Latent Dirichlet Allocation (LDA) in a practical application from a Requirements Engineering (RE) perspective in the analysis of AI training data. Given the scarcity of similar approaches in the surveyed literature, this work addresses a a contribution to help minimize this notable gap and try to contribute for future research focused on reconciling AI methods with ethical, privacy-preserving applications.

Keywords: Privacy Preserving Techniques, Differential Privacy, Semantic Text similarity, Rare Events, AI Agents, LLM, Topic Modeling, Latent Dirichlet Allocation, LDA, AI Training Data

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence.

BERT Bidirectional Encoder Representations from Transformers.

ChatGPT Chat Generative Pre-trained Transformer.

CRISP-DM Cross Industry Standard Process for Data Mining.

DLP Data Leak Prevention.

DM Data Mining.

DP Differential Privacy.

DR Dimensionality Reduction.

EC Exclusion Criteria.

EF Extraction Field.

FL Federated Learning.

GANs Generative Adversarial Networks.

GDPR General Data Protection Regulation.

GPT Generative Pre-trained Transformer.

IA Inteligência Artificial.

IC Inclusion Criteria.

IDF Inverse Document Frequency.

LaBSE Language-agnostic BERT Sentence Embedding.

LDA Latent Dirichlet Allocation.

LGPD Lei Geral de Proteção de Dados Pessoais.

LLM Large Language Model.

LLMs Large Language Models.

MCP Model Context Protocol.

ML Machine Learning.

MSE Mapeamento Sistemático de Estudos.

NER Named Entity Recognition.

NLP Natural Language Processing.

PICOC Population, Intervention, Comparison, Outcome, Context.

PII Personally Identifiable Information.

PLN Processamento de Linguagem Natural.

QAC Quality Assessment Criteria.

RE Requirements Engineering.

RF Random Forest.

RPART Recursive PARTitioning.

SMS Systematic Mapping Study.

TF Term Frequency.

Chapter 1

Introduction

This chapter outlines the contextual background of the study, articulates the underlying motivations for undertaking this research, and provides an overview of the proposed solution. Furthermore, it delineates the methodological approach adopted in the investigation. Finally, the chapter concludes with a description of the overall structure and organization of this document.

1.1 Contextualization

This work presents an overview based on studies on privacy-preserving in structured data and unstructured data [3] in text format, especially regarding differential privacy techniques [4, 5, 6, 7] and the risk of re-identification[8, 9] . In this context, contemplating the concepts of Natural Language Processing (NLP) [10], Artificial Intelligence (AI)[11], Agentic [12], Multi-Agent AIs [13], Large Language Models (LLMs) [14], (Model Context Protocol (MCP)) [15, 16] for Anonymization [17, 18, 19, 20, 21, 22][23] and for Rare Events Detection[24, 25, 26], with two case studies on the use of privacy-preserving techniques [27] considering data in text and data in structured format with the provision of software artifacts to enable the performance of similar experiments.

Stored data is increasingly becoming an object of interest, as raw material, for use in processes that seek to obtain wealth from the information contained in this data, according Schäfer et al. [1] these processes involve the creation of machine learning models, which tend to be more and as efficient as the data is more pure and complete, when integrating data mining techniques in process analysis and prediction. On the other hand, within this data there is information that can identify people and pose risks to their privacy when exposed in model creation processes. The use of such resources gives rise to a range of technical, social, and ultimately legal imperatives in machine learning models. In addition to core technical requirements, such as robustness, machine learning systems are

increasingly expected to demonstrate transparency and fairness (e.g., mitigating bias in decision-making processes) while ensuring the protection of data pertaining to all stakeholders, particularly data owners and model users [28].

Data protection laws were designed to safeguard personal data and privacy as regulatory frameworks[29]. As important marks we have General Data Protection Regulation (GDPR) [30] in Europe and the Lei Geral de Proteção de Dados Pessoais (LGPD) [31]. This happens at a time when our society is experiencing concerns about good commercial and technical practices and, especially, the need to comply with legal requirements regarding privacy protection. One rationale that represents a framework for delimiting limitations lies in the legal and ethical ramifications associated with textual data privacy. With the emergence of the GDPR and growing concerns over the privacy of sensitive content in unstructured data, such as images, videos, audio, and text, numerous studies have proposed differential privacy-based approaches to ensure rigorous privacy protections for such data [32, 33, 34, 35, 36, 37].

Regarding the relevance and academic significance of the present study, Saraiva and Soares [38] emphasized that solutions, techniques, and tools employed during the software development lifecycle have clear potential to support the protection and privacy of personal data. Nevertheless, they also underscore the existing gap in the literature concerning the extent to which artifacts produced by Information Security analysts can be effectively integrated into the software development process. In similar direction, Data Leak Prevention(DLP) systems are applied to monitor, detect, and control the flow of sensitive data to ensure compliance with security policies and regulatory requirements [39].

The characteristics of text data that may contain implicitly sensitive information can allow inferences to be made about an individual's identity, political beliefs, mental health status, and other attributes that are subject to protection under privacy regulations [39]. The present study follows this intuitive line of inquiry by identifying and defining its focus on textual data, including data stored in structured database fields and data embedded in longer, unstructured texts.

At this stage, the focus is primarily on textual data, however, the advanced domain of artificial intelligence cannot be disregard and it is important to recognize the significant interest in extracting value from these data sources and the wealth of textual data, which may contain sensitive or personally identifiable information (PII), presents a tension between being a valuable opportunity and a regulatory challenge. The rapidly advancing domain of artificial intelligence (AI), the strategic utilization of data has become a critical catalyst for productivity across a broad spectrum of industries. Nevertheless, a fundamental conflict has emerged between the imperative to uphold data privacy and the

increasing demand for secondary data use in AI-driven applications, particularly within fields such as deep learning. This conflict stems from the inherently divergent interests of key stakeholders: while data users require access to sensitive datasets for computational and analytical objectives, data owners are constrained by legal, ethical, and institutional responsibilities to safeguard the confidentiality of that data [40].

The challenge of balancing privacy with the great value that exists in fully preserved data, untouchable in its raw original form, is present in reconciling the seemingly contradictory objectives of data accessibility and data protection. There is a tension that presents a complex and pressing challenge in modern data governance, specifically, how to enable access to private data for operations such as cloud-based processing while simultaneously ensuring compliance with stringent data protection standards that preserve the rights and confidentiality of data subjects[40]. To achieve this desired goal, there are several privacy-preserving techniques[27, 41, 42, 43]. The selection of privacy-preserving techniques[44, 45, 46] may require domain-specific knowledge, as the business understanding phase constitutes a foundational step in a successful data mining (DM) project [1]. And also configurations to effectively handle the high dimensionality with dimensionality reduction (DR) or other techniques and contextual richness of unstructured data, and privacy-preserving mechanisms[6] are also sensitive to the unique properties of unstructured data. Considering multiple types of unstructured data is beyond the time and resources allocated for this study. Understanding the intuition that unstructured data has distinctions because it encompasses a diverse range of formats, such as images, audio, video, and text, and that privacy concerns regarding any type of unstructured data are broad enough for this study is essential to note that it is important to choose a narrow scope, which is to focus on unstructured data in texts.

Text data is a primary form of unstructured data generated in real-world contexts, characterized by the absence of a predefined data model. The primary concern in this context arises when such documents are subjected to text analytics operations, such as summarization, which involves extracting or generating a concise representation of the key elements within the document, by either third-party entities or in-house data scientists. These processes carry the potential risk of violating personal privacy [47].

This research delineates its scope to the examination of textual unstructured data, predicated upon its pervasive presence, inherent sensitivity, and unique vulnerabilities to privacy compromise. As we can see in the study by Yao et al. [48], unstructured data is a broad category of information that lacks a fixed schema and is often stored in its native format as the unstructured data model refers to data that has not been previously defined by the data model [48]. In contrast to structured data, which adheres to predetermined schema exemplified by relational databases, textual unstructured data encompasses

diverse manifestations of natural language communication, including electronic mail, conversational logs, social media contributions, and documentary texts. These data sources often contain personally identifiable information, confidential records, and other sensitive details. So susceptible to exploitation through unauthorized surveillance, identity theft, and data breaches.

Furthermore, advances in NLP and LLMs have exacerbated privacy concerns by allowing automated systems to infer latent personal information even from seemingly anonymized text corpora. There is an increasing reliance on machine learning and data-driven models, particularly in the field of NLP, has led to a growing demand for data sharing among organizations. However, as these datasets often contain personal information, they are subject to strict regulatory frameworks that necessitate anonymization prior to dissemination [49].

Textual data represent one of the principal forms of unstructured data generated in real-world contexts, distinguished by the absence of a predefined data schema. This category of data is predominantly encountered in digital documents and exhibits several defining characteristics, including the heterogeneity of file formats, its inherently unstructured composition, the embedding of domain-specific knowledge, the frequent presence of synonyms and orthographic variations, and a strong dependence on contextual cues for the accurate interpretation of linguistic elements. Moreover, the effective processing and analytical exploitation of textual data necessitate a rigorous preprocessing stage[47]. Moreover, unstructured textual data presents distinct computational challenges in preserving privacy, including the application of differential privacy [5] in text processing, secure multiparty computation for textual datasets, and privacy-preserving machine learning techniques tailored for NLP applications. In contrast to images and videos, which often employ perceptual anonymization techniques (e.g. blurring), textual data require sophisticated natural language anonymization strategies that strike a balance between utility and privacy. By focusing on unstructured textual data, this study seeks to contribute to a more nuanced and actionable understanding of privacy threats, risks, and mitigation strategies within the domain of textual communication and NLP-based data processing. This circumscription facilitates more granular analysis while ensuring the relevance of findings for practical applications in domains such as legal compliance, enterprise data governance, and AI-driven text analytics.

Privacy can be defined as an individual's right to control their self-disclosure, determining the extent to which they expose themselves to the external world [31]. This concept is closely linked to key data protection principles, including data minimization, anonymity, and informed consent, which collectively aim to safeguard personal information while ensuring individuals retain agency over their data [50]. Certain benefits of

information technology can only be realized through the collection and analysis of data, which may at times include confidential or sensitive information[27]. A significant concern arises from the substantial volume of sensitive personal information embedded in data sources. Some attributes are inherently unique to individuals and, once disclosed, pose a permanent risk to personal privacy. As a result, ensuring the protection of sensitive content in unstructured data before it is shared with untrusted parties represents a critical challenge in contemporary data privacy research[32].

Data mining is the process of identifying patterns or constructing models based on observed data[51]. Therefore, we can infer that since privacy techniques alter the original data before using it to create models, the models are far from the true relationship with the content of the initial data and are therefore less efficient. In statistical learning, privacy has become an increasingly critical concern, particularly when handling large volumes of confidential data within an organization or sharing such data externally. Ensuring that researchers can assess data utility while preventing the leakage of sensitive information or compromising the privacy of individual records is a fundamental challenge. As data-driven decision-making and machine learning applications continue to expand, striking a balance between data utility and privacy preservation remains a key research focus[52]. The more we protect information, the more it loses its original value as a source of patterns[52]. The richness of processed and polished data is directly related to its purity. The more we obscure the original data by scrambling, encoding, and encrypting, we move from maximum utility to minimum. For example, data encrypted with a strong algorithm has maximum protection, one might say privacy, but on the other hand, if it is not decrypted yet or not decipherable, it has zero value for use in machine learning models.

Sometimes the pseudoanonymization [53] could be the goal, for example, in scenarios where the re-identification is needed after the use of machine learning models. In other way, if anonymization [18] is the final goal, achieving complete anonymization is inherently challenging. In narrative texts that describe life events, even if the anonymization process is fully successful in removing direct identifiers, the underlying facts often remain intact. When such facts represent rare or unique events, there is a heightened risk of re-identification, as it may be possible to infer personal information by linking these uncommon details to specific individuals. In this sense, when dealing with risks of identification and re-identification, with the consequent discovery or inference of personal information about an individual or a small group of people. In this way we have the concepts involving Differential Privacy presented originally as an articulation of a desideratum for statistical databases: nothing about an individual should be learnable from the database that cannot be learned without access to the database, but that this is not possible because, in terms of semantic security, it cannot be achieved. And that, counter-

intuitively, a variant of the result threatens the privacy even of someone who is not in the database. This scenario suggests a new measure, differential privacy, which intuitively captures the increased risk to someone's privacy incurred by participating in a database. The techniques developed can achieve any desired level of privacy with this measure. In many cases, extremely precise information from the database can be provided, ensuring high levels of privacy[5].

Then rare events detection is another part of the study object here. Work with rare events detection begin with the understanding of how much a new data, a new text, with its facts is usual or rare, common or an anomaly. Does it have characteristics of a normality or outlier? That's an relevant question. One way of mensuration is the use of embedding to represent words as vectors and calculate their representative distance and do the same with groups of words, sentences and the whole text. Mikolov et al.[54] introduced a foundational approach to learning distributed word representations through neural network-based models with distributed representations of words in a vector space enhance the performance of learning algorithms in natural language processing tasks by clustering similar words, in a semantic comparison, thereby enabling models to capture linguistic patterns and generalize more effectively.

Particularly regard to rare events and the implications of data memorization in machine learning models, while such models are designed to extract generalizable patterns from training data, they are not intended to memorize rare data points or outliers. The memorization of these rare instances can undermine both the utility and the privacy of the model. Specifically, when a model learns disproportionately from a small subset of individuals, it increases the risk of information leakage, as such data can be more easily traced back to its origin. For instance, it is reasonable to expect a language model like ChatGPT to recall publicly available information, such as the fictional address of Harry Potter, but not to retain or reveal private, hard-to-find personal information, such as an individual reader's address [49]. The same authors further note that the most widely adopted approach for mitigating privacy risks in machine learning models is Differential Privacy (DP).

The business and the data related must be understood, including where is the data, before to choose and apply techniques to protect it. Having this purpose, especially in the case study, this study follow Cross Industry Standard Process for Data Mining (CRISP-DM)[1, 2], that is an industry-independent process model for data mining, with six phases: Business Understanding, Data understanding, Data preparation, Modeling Evaluation and Deployment.

1.2 Research Problem

Studies have shown a regulatory evolution about how to safeguard personal data and privacy as regulatory frameworks[29], a pressing challenge has emerged around the privacy of unstructured data that are more difficult to anonymize[17, 18, 19, 20, 21, 22][23] than structured data formats due to their high dimensionality and semantic richness[55]. Two landmark legislative frameworks are the GDPR[30] in Europe and the Brazilian General Data Protection Law (LGPD)[31].

When discussing the protection of sensitive information within systems, to prevent data leakage, DLP systems are designed to prevent authorized users within an enterprise network from either ex filtrating sensitive information beyond the network boundaries or introducing unauthorized sensitive data into the system and that a effective DLP solution typically comprises multiple tools or components, which are generally categorized into host-based and network-based DLP mechanisms[39].

In the context of an institution with thousands of records and seeking to enrich its data, the aim is to leverage the benefits of Data Science and LLMs [14] in AI [11] and Agent Techniques [13]. To achieve this, third-party services and companies are required to create and process the models and implement the technologies. On the other hand, if part of this data is confidential, data privacy must be guaranteed. So there is a point of attention, which is to protect the privacy of data before it is made available to third parties for model creation work.

Textual data eventually contains implicitly sensitive information, identifiers and quasi-identifiers which can enable the inference of personal attributes[39]. Sometimes texts contain sensitive information about individuals[40]. However, effective AI models require high-quality data, meaning anonymization, so the privacy techniques could reduce model performance[27].

In light of prior research addressing the risks of data leakage, the necessity of implementing protective techniques, and the importance of ensuring compliance with regulatory frameworks, this study is guided by the objective of to explore techniques and strategies that balance the trade-off between privacy preservation and the usefulness of data for AI applications. And then, addresses the challenge of balancing privacy protection[27] and data utility in datasets used for AI applications and we purpose a process using anonymization techniques[17, 18, 19, 20, 21, 22][23] and natural language processing[10], which includes concepts of differential privacy[32, 33, 34, 35, 36, 37] and detection of rare events[24, 25, 26], contributes to the privacy of people in the preparation of the texts data before they are used to create artificial intelligence models[12, 13, 14, 15, 16].

1.3 Aims and Objectives

The main objective was to identify techniques for the privacy-preserving handling of text data prior to its utilization in the development and training of artificial intelligence models. To achieve this purpose we undertake a examination of privacy-preserving techniques with the objective of exploring their applicability to structured data and textual content.

The practical side of this goal is to clear the understanding of how to protect privacy in texts that are raw materials of use in AI models before them are used for this purpose. For a better understanding, we can say that it is the preparation phase before their use in AI models. We looked at studies that considered the balance of maintaining a high level of data utility while ensuring the privacy of those involved leveraging classic privacy-preserving techniques, data science techniques, including vectorization, and also AI and LLMs and agent-based models to safeguard sensitive information. Additionally, we analyze the potential privacy risks associated with the presence of rare events, emphasizing how such events may facilitate re-identification due to their uniquely identifiable characteristics as rare facts.

The specific objectives of this study were articulated as follows:

- To propose the application of privacy-preserving and similarity techniques in texts leveraging the principles and risks of re-identification in the domain of differential privacy[32, 33, 34, 35, 36, 37].
- To develop software artifacts as a means to implement the use of the techniques and concepts like the use of vector databases, data science and machine learning techniques, leveraging LLM and AI agents, for rare events detection and text similarity.

1.4 Expected Results

The primary contribution of this work is the selection of a set of privacy-preserving techniques tailored for the anonymization of both structured and textual data designed to safeguard personal data embedded in textual content prior to its use in the development of AI models. These guidelines are intended to support professionals responsible for ensuring compliance with privacy legislation, ethical principles, and technical data protection requirements.

In addition to privacy preservation techniques identification, a process was proposed for apply techniques and concepts in a case study.

This research aimed to advance privacy-preserving methodologies by selecting, implementing, and evaluating effective techniques in the context of texts. The combination of

theoretical analysis, empirical experimentation, and AI-driven techniques can contribute to advancing knowledge in privacy-preserving and text similarity data analytics. The study's findings are expected to offer valuable insights for policymakers, data custodians, and researchers working on privacy-enhancing and tex similarity technologies in the context of high-risk institutions with regard to the holding of sensitive personal data.

1.5 Methodology

This research followed a structured divided into two distinct phases, ensuring a approach to achieve the defined objectives. The second phase builds on the previous one, incorporating theoretical and experimental components to select and purpose a implementation process of techniques and concepts from the literature review.

The first phase involved a review of existing privacy-preserving techniques, differential privacy models, and rare event detection methodologies. This step was conducted by performing a Systematic Mapping Study (SMS), which provided the foundational knowledge base as a theoretical basis for the study. The literature review focused on key works in the domain of privacy-enhancing technologies, re-identification risks associated with quasi-identifiers, and the effectiveness of AI models in privacy protection and rare event detection. In addition, research on the applicability of large language models (LLMs) in data anonymization were explored. The output of this phase was a synthesis of existing knowledge, tried to identify gaps and best practices relevant to the objectives of the study and this served as input for phase 2 for techniques and concepts selection.

After the literature review and based on it, the second phase was materialized on selecting techniques that were applied in a study case. It was a process that considered differential privacy concerns [5, 7] to build the process. The case study examined the role of Agent AI-driven methods, particularly LLMs, in automating privacy-preserving techniques and rare event detection. The review of Agent AI applications in privacy research and recent advances in machine learning-driven processes informed this analysis. Experiments were conducted using LLM-based tools to examine the capabilities and limitations of LLMs in supporting the implementation of techniques to improve privacy preservation and about text similarity. This phase also involved rare event detection concepts, particularly with regard to re-identification risks in anonymized datasets. Rare events, such as outlier incidents, can serve as quasi-identifiers that can compromise anonymity despite traditional anonymization techniques. The case study also involved the development of software artifacts based on the previous concepts to apply the techniques and concepts from phase one.

The application of techniques and concepts identified in the literature review began with a focus on privacy-preserving methods. However, their scope extended beyond that domain. Most of the studies reviewed in the SMS emphasized methods related to text similarity rather than being exclusively centered on privacy. Consequently, it was not a surprise that the case study, although initially designed with specific goals related to privacy preservation, would also yield relevant and applicable results in areas involving inference from text similarity analysis. This outcome aligns with the broader methodological trends observed in the literature and reflects the multifaceted utility of these techniques.

1.6 Publications

Publication originated from this research:

1. Privacy Preservation in Textual Data: A Systematic Mapping Study on Differential Privacy and Semantic Similarity (CIbSE 2026)

1.7 Data Availability

The software artifacts created for conducting the experiments in the case study are available at https://github.com/Daniel-Lim-Apo/dissertation-experiment-01, at https://github.com/Daniel-Lim-Apo/Topic-Modeling-Privacy-for-AI-Training and https://zenodo.org/records/19261640.

The remaining supporting data for this work cannot be made available due to its confidential nature.

1.8 Manuscript Organization

This manuscript comprises six chapters, including the present one. The structure of the dissertation follows the progression of the research process, with each chapter corresponding to a specific phase. Below, a summary is provided for each chapter, indicating the phase it represents and an overview of its substantive content.

- Chapter 2 — Background: An overview of privacy-preserving in structured data and unstructured data in text format, especially regarding differential privacy techniques, introduces the concept of NLP, AI Agentic and Multi-Agent LLMs, Anonymization and Rare Events Detection.

- Chapter 3 — Literature Review (Phase 1): A review of existing privacy-preserving techniques, differential privacy models, and rare event detection methodologies establishes the theoretical foundation for this study; — Privacy Techniques (Phase 2): Based on the literature review, the second phase focuses on selecting privacy-preserving techniques that can be applied to the case under study. This involves the anonymization of direct identifiers and quasi-identifiers in texts. The study will consider different privacy issues;
- Chapter 4 — Case Study: A real case implementation; Leveraging AI Agents LLM: The study examined the role of Agent AI-based methods, particularly LLMs, in automating privacy-preserving techniques and detecting rare events and working with text similarity;
- Chapter 5 — Case Study: Latent Dirichlet Allocation (LDA) to extract latent topics from raw textual data to identify privacy risks before data are incorporated into AI systems.
- Chapter 6 — Conclusion: Provides a summary of the work conducted and the results obtained.

Chapter 2

Background

This chapter provides an overview of privacy-preserving in structured data and unstructured data in text format and its regulatory concerns, especially regarding differential privacy techniques, introduces the concept of NLP, AI Agentic and Multi-Agent LLMs, Anonimization and discusses Rare Events Detection.

2.1 Preserving privacy in texts

The abundance of text data has made presents a challenge for artificial intelligence[56]: how to leverage sensitive personal information without compromising privacy[57]. As AI models become more advanced, balancing data utility with privacy protection is important[18]. This chapter establishes the foundation for our research by reviewing the concepts and techniques for privacy-preserving text analysis.

To support a investigation of privacy-preserving approaches in AI[50, 58], this chapter is structured around three thematic areas aligned with the research questions: foundational concepts of privacy, including techniques for anonymization[18, 17, 20, 19] and protection of textual data[21, 22], and specialized methods for semantic similarity[59, 60, 61] and rare event detection[25, 26, 24] under privacy constraints.

The concept of privacy has been subject to extensive theoretical debate and regulatory evolution, particularly in the context of personal data processing. Foundational theories such as contextual integrity[62, 63] emphasize the importance of preserving informational norms, while legal frameworks and regulatory compliance themes like the GDPR[30, 64, 57] and LGPD[31, 38] provide formal definitions of personal data and obligations regarding its treatment. In the domain of textual data, privacy concerns are amplified by the unstructured nature of the content[32, 65, 58] and the risk of re-identification[9] through seemingly benign text fragments.

The idea that user privacy can be compromised under various circumstances is a central insight derived from the study by Cauvery and Kulkarni [66]. Such violations, for them, may occur when datasets originating from multiple sources are aggregated, thereby increasing the risk of unintended information disclosure. Furthermore, privacy breaches may arise when personal data collected for a specific purpose are subsequently utilized for an entirely different objective, violating the principle of purpose limitation. In addition, sensitive information stored and processed in inadequately secured environments remains vulnerable to unauthorized access and data leakage during both storage and processing phases. The authors therefore emphasize the necessity of redefining privacy-preserving mechanisms to effectively address these emerging challenges.

Research and studies in recent years have examined data leakage risks, emphasizing the need to apply robust privacy-preserving techniques in natural language processing to align processing interests with evolving regulatory and ethical frameworks, the well-cited trade-off [67, 21, 68, 49] between privacy protection and data utility [69].

This study addressed this challenge by seeking to gather from academia which anonymization techniques are used with NLP[70, 71, 72, 73, 74] methods to enhance the privacy of textual data prior to its use in the development of AI models. In this regard, it is interesting to consider the concepts of anonymization and pseudo-anonymization.

The approach is also grounded in the theoretical foundations of differential privacy [75, 76], a formal framework that quantifies privacy loss and aims to protect it even in the presence of auxiliary information. Recent research and technical advances[32, 33, 34] demonstrate the applicability of differential privacy to textual and unstructured data. Techniques from the field of rare event detection[24, 25], including innovative classification and semantic outlier detection [26] are integrated to enhance privacy guarantees.

Differential privacy, which saw its seed grow from the work of Dwork [4] and other authors studies[32, 33, 34, 35, 36, 37], highlights the interest in identifying what is common and what is rare. A text with fully anonymized[17, 18, 19, 20] data can still be used to identify individuals if the facts of life present in the text are rare, rare events. Therefore, attention is being paid to detecting and protecting these rare events in textual corpora, which pose re-identification risks. To identify what is rare, one approach is to identify what is similar in a given domain, more specifically in a dataset, and how to group it, clustering it into similar sets. This was demonstrated concretely through an experiment conducted during the case study phase. From this process, we have two outcomes: identifying what is rare and also identifying what is similar. These two outcomes have their benefits, both from a privacy preservation perspective and in extracting richness from similar data, thus serving semantic understanding, pattern identification, and even text prediction. This has been fruitfully explored in recent studies, as we saw during the research and will address

in the discussion section of this study.

By addressing these components, privacy protection with anonymization[21, 22][23] and other techniques in an NLP pipeline, considering differential privacy[75, 72, 77] and detection of rare and similar textual components, this study sought to contribute to the literature on data preparation pipelines that preserve privacy in data preparation, for disclosure in compliance with transparency requirements, and, especially, as texts to be used as raw material for generating AI models or for use in AI pipelines. Thus, there is alignment with ethical and responsible AI for context-sensitive content.

The integrated detection of rare events[25, 26], along with the calculation of corresponding rarity metrics[24], serves to inform data managers about the presence and extent of re-identification risk[9]. This enables informed decision-making regarding the handling, retention, or transformation of potentially sensitive data elements. The expected outcomes included a formalized understanding of how rare events are relevant in strategies to mitigate the privacy risks, leveraging the differential privacy concepts and the wealth of text semantic similarity detection.

A survey on rare event identification [25, 26, 24] offered a state-of-the-art evaluation of existing detection methodologies, with a particular focus on those applicable to the specific business context. The study introduced ideas for quantifying event rarity, incorporating approaches from outlier analysis and anomaly detection models. The expected outcome includes a formalized definition of event rarity within the given database context, along with insights designed to support decision-makers in determining appropriate responses to process or even don't process the identified rare events. Demonstrate whether vectorized databases offer significant advantages for rare event detection.

2.2 Related Works

Our study concerned about what Privacy-Preserving Techniques are applied in textual data analysis, in this track we found the study of Nethravathi et al. [78] in an approach that encapsulates various techniques of text-processing, keyphrase extraction, co-occurrence analysis, ontology construction and query analysis.

Another study, with a similar path in the work of Fei et al. [74], that centers on the protection of privacy in health records within the context of big data analytics and introduces a privacy-preserving framework grounded in NLP methodologies. The authors employ techniques such as word embeddings and transformer-based architectures to develop a system capable of analyzing and safeguarding sensitive medical information. Leveraging deep learning and self-attention mechanisms, the proposed system enables the automated identification and encryption of sensitive textual data in medical records.

This approach aims to achieve a balance between data privacy protection and and medical information research.

Concerning the identification and comparison of text similarity techniques, Bernard et al. [79] conducted a systematic comparison of term-based text similarity measures and machine learning classifiers to identify the most effective combinations for improving the accuracy and efficiency of research paper recommender systems.

The work of Sitikhu et al.[80] focuses on improving the measurement of semantic similarity between short texts. The authors aim to ensure that these similarity measures align closely with human interpretability, a key requirement for applications such as information retrieval, text classification, and natural language understanding. Their work presented techniques as Euclidean distance, Cosine distance, Jensen Shannon Distance, Word Mover distance as being distance metrics used in the computation of text similarity. And for generate features from the documents or corpus: Tf-idf vectors, Word Embeddings and the methods cosine similarity with tf-idf vectors, Cosine similarity with word2vec vectors and Soft cosine similarity with word2vec vectors.

In recent years, the efficient management of high-dimensional vector data has become an increasingly critical challenge in data science and AI applications. This growing demand is driven by the widespread adoption of unstructured data and ML techniques, wherein ML models frequently convert unstructured data into feature vectors for analytical tasks, such as product recommendation. However, existing systems and algorithms for vector data management exhibit two primary limitations. First, they suffer from significant performance inefficiencies when processing large-scale and dynamic vector data. Second, they offer limited functionality, failing to meet the diverse and evolving requirements of modern applications[81].

This review will highlight state-of-the-art methodologies, identify key trends, and assess the extent to which LLMs and other AI agents contribute to privacy-preserving frameworks. The results will provide a structured synthesis of the impact of AI in this domain, outlining best practices and existing challenges. Agentic AI, an emerging paradigm in artificial intelligence, refers to autonomous systems designed to pursue complex goals with minimal human intervention. Unlike traditional AI, which relies on structured instructions and close supervision, Agentic AI demonstrates adaptability, advanced decision-making capabilities, and self-sufficiency, enabling it to operate dynamically in evolving environments[12].

In addition to agents, there are proposed multi-agent solutions. Tang[82] in his empirical work presented experimental results to demonstrate that the outcomes of multi-agent dynamic interactive learning surpass those of individual site learning in his study experiments. The primary advantage of agentic and multi-agent systems lies in their ability to

decompose complex tasks, enabling goal achievement through the collaborative actions of multiple agents. This approach enhances the system's flexibility and adaptability while also improving its capacity for generalization across diverse scenarios[13].

Large pretrained language models LLMs are particularly valuable for languages with limited annotated resources but abundant unlabeled data, such as Brazilian Portuguese. By leveraging self-supervised learning on extensive unstructured text, these models can effectively capture linguistic patterns and enhance performance in various NLP tasks, even in low-resource settings[10]. An analysis of the ability of LLMs to perform or enhance privacy-preserving tasks, such as detecting personally identifiable information, generating synthetic data, and applying differential privacy techniques.

In this study, we emphasize the necessity of safeguarding privacy at all stages of the use of an AI model, including the textual data employed as raw material for training. In this regard, it is important to highlight the study of Behnia et al.[76] with a discussion where, when discussing large-scale pre-trained language models (LLMs), adversarial attacks can be used to extract or even reconstruct exact samples from their training data, thereby exposing personally identifiable information (PII). Such vulnerabilities raise serious concerns about the privacy of LLMs. Differential Privacy (DP) provides a rigorous theoretical framework to mitigate these risks by introducing carefully calibrated noise during the training or fine-tuning process. Addressing this challenge, Behnia et al. propose EW-Tune, a differential privacy framework specifically designed for the fine-tuning of LLMs.

The study undertaken by Chakrabarti et al. [83] does not focus directly on privacy threats and risks. Rather, it encompasses a spectrum of commercial concerns, including liability, indemnity, confidentiality, and other related forms of organizational risk. Nevertheless, the work is relevant as related literature, particularly due to its methodological contribution. The authors propose a framework, termed the Risk-o-Meter, which leverages ML and NLP techniques to identify risk-prone textual segments and map them to predefined risk categories.

Another related work, we should mention lead us that sensitive information detection (SID) as a subpart of data leak detection (DLD) that deals with the automatic identification of sensitive information is a concept present in the study conducted by Gambarelli et al. [84]. The work also contributes to improving data loss prevention (DLP) systems to avoid data breaches, presenting a way to train, classify and perform the classification of sensitive text and addresses the challenge of identifying complex personal information in unstructured text.

The safeguarding of privacy has gained growing significance within the field of NLP. One direction in this area is anonymization, which involves removing identifying infor-

mation from the text corpus. More recently, obfuscation, which replaces any sensitive information with a different substitute of the same type, has been investigated[73].

The study conducted by Martinelli et al.[85] acknowledges that Artificial Intelligence (AI) has introduced novel methods and solutions across various domains of application. It further recognizes that the field of Privacy and Data Protection constitutes a frontier in legal, regulatory, academic, and technological developments. In their research, the authors employed both supervised and unsupervised machine learning techniques on documents from the legal and medical domains to produce annotated corpora in a semi-automated manner with a pipeline with a Knowledge Extraction(KE) leveraging PoS-Tagging and Dependency Parsing, Named Entity Recognition (NER), Context Window, Transfer Learning and Word Embeddings Computing, Topics Extraction and then they made a Knowledge Fine Tuning labeling Named Entities, categorizing "sensitive semantic categories". In this approach, they used the NLP and spaCy and the human expertise was required only in the final stage of the process, to validate and refine the data automatically generated, for the purpose of detecting and classifying sensitive privacy-related information in textual documents.

About the goal of purposing a process, we can cite the study conducted by Saeed et al.[86] that presents a well-articulated methodology aimed at addressing the challenges posed by the heterogeneity of unstructured textual data. The proposed approach offers a cost-effective, language-structure independent framework that leverages multiple techniques to effectively identify and extract semantically relevant segments within a corpus. Notably, the methodology is highly detailed, conceptually rich, and demonstrates a clear and systematic exposition of its underlying processes.

Vinaykumar et al[87]. propose novel dimensionality reduction techniques for text data mining that leverage feature similarity measures to cluster features and construct a transformation matrix. This matrix is then used to project high-dimensional text data into a lower-dimensional space, enhancing the efficiency of text clustering and classification tasks.

Lai et al[88]. proposed a sentence vector similarity method that employs a weighted fusion approach based on the FastText model, integrating both word- and sentence-level features. Similarity is calculated using cosine similarity between weighted sentence vectors, as well as a word-level W-WRD algorithm. The method demonstrates strong performance in terms of accuracy, robustness, and transferability across different datasets.

Huang et al[89]. presented a short text similarity model that integrates both semantic content and word order information. The model combines an adjacency-aware semantic module, implemented through deep convolutional neural networks (CNNs), with a word order module that incorporates external knowledge bases and pointwise mutual infor-

mation. Experimental results on the MRPC dataset indicate that the proposed model outperforms existing approaches.

Zhang et al[90]. introduce an enhanced text similarity algorithm that extends traditional cosine similarity by incorporating both vector direction and the rate of change across dimensions. This approach aims to improve the accuracy and efficiency of similarity measurements.

Acharya et al[12]. conducted a comprehensive survey of Agentic AI, emphasizing its capabilities in autonomous goal pursuit, adaptability, and decision-making within dynamic environments. The study examines foundational methodologies, diverse applications across sectors, and associated ethical challenges, ultimately proposing a framework for the responsible integration of Agentic AI into society .

Chandrasekaran[91] investigated how the SmartTaskAgent and CollaborativeAI frameworks enhance efficiency and scalability in the development of multimodal intelligent agents through the integration of large language models. The study addresses key challenges, including role assignment, prompt reliability, hallucination reduction, and workflow scalability, demonstrating the frameworks' potential for rapid prototyping and deployment across various industries.

Z. Duan and J. Wang [13] investigated the integration of LangGraph and CrewAI to enhance multi-agent systems, with a focus on improving information flow, task collaboration, and overall system performance. The study presents architectural designs aimed at enabling precise agent control and explores mechanisms for intelligent task allocation, providing insights to the advancement of large model-based agent technologies.

2.3 Chapter Summary

This chapter established a foundation for the concepts of preserving privacy in textual data for artificial intelligence applications. This issue was contextualized within current legal frameworks, such as the GDPR and the General Data Protection Law (LGPD), while also outlining specific characteristics of unstructured text, including identifiers, quasi-identifiers, rare events, and similar texts detection.

Chapter 3

Systematic Literature Review

This chapter presents a review of the relevant literature. To achieve this purpose, a Systematic Mapping Study (SMS) was conducted in accordance with the established protocol proposed by Kitchenham and Charters [92], with the aim of uncovering prevailing practices and techniques in privacy-preserving and use on text analyses. The SMS methodology involves a structured and replicable process for identifying, evaluating, and synthesizing existing research relevant to a particular area of inquiry or research question. As delineated by Kitchenham and Charters [92], the SMS process comprises three main phases:

1. *Planning Phase*: This phase involves establishing the necessity of the review, formulating research objectives, and developing a review protocol. The protocol includes as components: the formulation of research questions, the construction of a appropriate search string, the definition of inclusion and exclusion criteria for study selection, the specification of data extraction elements, and the design of a quality assessment checklist.

2. *Conducting Phase*: This phase entails the implementation of the review protocol developed during the planning phase. It includes the identification and selection of primary studies, followed by the extraction and synthesis of relevant data in accordance with the predefined criteria and procedures.

3. *Reporting Phase*: In the final phase, the results of the review are documented and presented in the form of a research manuscript, providing a transparent and account of the findings and their implications.

3.1 Research Questions

The Systematic Mapping Study (SMS) was undertaken with the objective of uncovering contemporary and state-of-the-art practices in privacy-preserving techniques[27] applied to textual data. It is important to emphasize that the scope of this review is confined exclusively to text-based data; other forms of unstructured data, such as images, audio, or video, fall outside the purview of this study. The research questions guiding the review are outlined in the problem of how to anonymize unstructured data in texts for use in artificial intelligence models, maintaining a high level of data utility while ensuring the privacy of those involved.

The research objective was to examine how to anonymize or protect personal data in structured and unstructured data in texts for use in artificial intelligence models, maintaining a high level of data utility while ensuring the privacy of those involved leveraging classic privacy-preserving techniques, data science techniques, including vectorization, and also AI and LLMs and agent-based models to safeguard sensitive information in preparation phase before the use in AI models and which techniques are employed for semantic similarity and rare events detection in text data under differential privacy constraints. To address it in this study is relevant to identify what privacy-preserving techniques have been applied in textual data analysis, including those that leverage typical Data Science, LLM and Agent-Based AI Techniques. Furthermore, what techniques are employed for semantic similar and rare events detection in text data under differential privacy constraints are considered in a context of protecting personal information in high-risk domains.

This goal was decomposed into three interrelated components: 1) which addresses the Privacy-Preserving Techniques; 2) which focuses on Data Science, LLM and Agent-Based AI Techniques; and 3) which examines the semantic similarity and rare events detection. This structured analysis supports the exploration of the research problem with analytical clarity throughout the study. Those components are realized in three research questions:

RQ.1: **What Privacy-Preserving Techniques are applied in textual data analysis?**

This question investigates the intersection between data privacy and textual data. We looked at privacy concerns and techniques to preserve privacy in texts.

RQ.2: **Which Data Science, LLM and Agent-Based AI Techniques are used to implement privacy-preserving mechanisms in text analysis?**

This research question examines how three strands of computational approaches as data science, LLMs, and agent-based AI are leveraged to design and implement privacy-preserving mechanisms in text analysis.

RQ.3: **What techniques are employed for semantic similarity and rare events detection in text data considering differential privacy?**
This question investigates how privacy-preserving methods, especially Differential Privacy (DP), are integrated into text analysis tasks that require semantic sensitivity comparing meanings of sentences, documents, or embeddings, identifying infrequent but significant textual patterns.

3.2 Search String

Petticrew and Roberts [93] suggested that research questions should be formulated with focus on five elements known as PICOC. This methodological framework, PICOC, is a commonly used in systematic literature reviews to structure research questions and develop search strategies by clearly defining five main elements: the Population (the subject or data under study), the Intervention (the technique or approach applied), the Comparison (an alternative to the intervention, if applicable), the Outcome (the effect or outcome being measured), and the Context (the conditions, domain, or environment in which the study is situated).

As outlined in Table 3.1 the population focuses on textual datasets and data subjects whose information requires privacy protection, particularly when such text contains sensitive, identifiable, or event-related content. In this way, the population was set as textual datasets with facts, events, containing sensitive content. Relevant keywords are provided to facilitate reproducibility and the retrieval of literature, encompassing forms of unstructured text, sensitive documents, and privacy-sensitive corpora.

Within the PICOC framework, the Intervention component specifies the techniques, methods, and technological approaches applied to address the research questions. As outlined in Table 3.1, the interventions considered in this review include the application of privacy-preserving techniques, such as anonymization, pseudonymization, and differential privacy, combined with Data Science, LLMs, and agent-based AI approaches for semantic similarity analysis and rarity detection in textual datasets. The table further details a set of keywords and corresponding search strings to support a systematic and reproducible literature retrieval, methods from privacy-enhancing technologies to advanced natural language processing and anomaly detection techniques.

In the PICOC methodology, the Comparison element identifies alternative interventions or baselines against which the primary approach is evaluated. For this SMS, as shown in Table 3.1, no explicit comparison criteria were defined. This decision reflects the focus on surveying and synthesizing privacy-preserving techniques and AI-driven analytical

methods without benchmarking them against specific alternative approaches, making the Comparison component not applicable in this context.

The Outcome element, in the PICOC framework, defines the expected results or measurable effects of the interventions under study. As presented in Table 3.1, this review targets outcomes that enhance privacy in textual data analysis while maintaining or improving analytical utility. Key objectives include the protection of sensitive information, preservation of semantic richness, detection of rare or distinctive textual patterns, and a balance between privacy and data usability. The table also specifies associated keywords to serve the search string building supporting retrieval of relevant studies addressing interpretability, utility–privacy trade-offs, and quality preservation in privacy-preserving text processing.

Within the PICOC framework, the Context element specifies the settings, domains, or application environments in which the studied interventions and outcomes are relevant. As shown in Table 3.1, this review focuses on sensitive or high-risk domains where textual data analysis demands both high interpretive value and privacy safeguards. These contexts include AI-driven natural language processing in regulated sectors, legal and compliance analytics, and environments addressing re-identification risks under differential privacy principles. The table also provides targeted keywords to guide the retrieval of studies situated in domains emphasizing responsible, ethical AI applications.

3.2.1 Search String

To achieve a reproducible and transparent retrieval process, a search string was constructed by integrating all elements of the PICOC framework defined in this review. The string combines Population, Intervention, Outcome, and Context components using operators to capture the full range of relevant literature. This formulation enables systematic querying across multiple digital libraries and indexing services, ensuring coverage of studies on privacy-preserving textual data analysis, AI-driven semantic similarity methods, privacy–utility trade-offs, and domain-specific applications in sensitive or regulated contexts.

We began by identifying the primary keywords for our PICOC framework: textual data, sensitive text, privacy-preserving techniques, semantic similarity, rarity detection, privacy in textual data analysis, protection of sensitive content, detection of semantic structure and rare or distinctive patterns,natural language processing, ethical AI, information retrieval. Subsequently, we conducted exploratory searches using related terms that we had initially defined, with the aim of evaluating the relevance of the retrieved results and identifying any pertinent terms that had not been previously considered. Fol-

Table 3.1: PICOC terms

PICOC	Keywords	Related Words
Population	textual data, sensitive text	unstructured text, natural language data, corpora, text, private data, confidential document, textual dataset, personal text data, privacy-sensitive textual corpora
Intervention	privacy-preserving techniques, semantic similarity, rarity detection	privacy protection, anonymization, pseudonymization, differential privacy, privacy-enhancing, re-identification, AI, data science, data analytics, machine learning, artificial intelligence, AI model, GPT, BERT, agent, predictive modeling, statistical modeling, natural language processing, NLP, language model, transformer model, sentence similarity, document similarity, semantic matching, text similarity, cosine similarity, embedding similarity, vector similarity, semantic distance, similarity metrics, lexical similarity, similarity detection rare word, low-frequency term, outlier, anomaly, rare entity, statistical rarity, frequency analysis
Comparison	*Not applicable*	*Not applicable*
Outcome	privacy in textual data analysis, protection of sensitive content, detection of semantic structure and rare or distinctive patterns	interpretability, analytical utility, semantic richness, text clustering, topic modeling, document classification, privacy protection, risk mitigation, data anonymization, balance between utility and privacy, privacy-utility trade-off, knowledge extraction, meaning preservation
Context	natural language processing, ethical AI, information retrieval	privacy in AI, responsible AI, legal NLP, AI agents, autonomous agents, conversational AI, large language models, LLM, foundation models, data science.

lowing several iterative refinements of this process, we arrived at the final, adjusted generic search string:

("textual data" OR "unstructured text" OR "natural language data" OR corpora OR "sensitive text" OR "private data" OR "confidential document" OR text OR "textual dataset" OR "personal text data" OR "privacy-sensitive textual corpora") ***AND*** *("privacy-preserving techniques" OR "privacy protection" OR "anonymization" OR "pseudonymization" OR "differential privacy" OR "privacy-enhancing" OR "re-identification" OR "semantic similarity" OR "sentence similarity" OR "document similarity" OR "semantic matching" OR "text similarity" OR "cosine similarity" OR "embedding similarity" OR "vector similarity" OR "semantic distance" OR "similarity metrics" OR "lexical similarity" OR "similarity detection")* ***AND*** *("interpretability" OR "analytical utility" OR "semantic richness" OR "text clustering" OR "topic modeling" OR "document classification" OR "privacy protection" OR "risk mitigation" OR "data anonymization" OR "balance between utility and privacy" OR "privacy-utility trade-off" OR "knowledge extraction" OR "meaning preservation")* ***AND*** *("Natural language processing" OR NLP OR "Information retrieval" OR "Privacy in AI" OR "Responsible AI" OR "Ethical AI" OR "Legal NLP" OR "AI agents" OR "Autonomous agents" OR "Conversational AI" OR "Large language models" OR "LLM" OR "Foundation models" OR "Data science")*

The digital databases selected for executing the search string were the ACM Digital Library[94], IEEE Xplore[95], Scopus[96], and Web of Science[97]. The search strategy was applied across four databases, a decision that was guided by the criteria proposed by Kitchenham et al. [98], who identified them as essential sources for conducting systematic literature reviews in the field of Software Engineering, particularly for their practice in indexing of high-quality journals and conference proceedings, and for supporting the execution of the complete search string without modification. Merrouni et al.[99] informed that ACM, IEEE and Scopus host many of the most reputable computer science papers. This present study was in part Data Science and also in the field of Software Engineering. The generic search string was adapted for each digital library and the access to the databases was provided via the CAPES Periodicals Portal, using institutional credentials from the University of Brasília.

3.3 Selection Criteria

Inclusion Criteria (IC)

Studies were included in the review if they meet the criteria in Table 3.2:

Table 3.2: Inclusion Criteria (IC)

Class	IC	Criteria
Topical relevance	IC-1	Focus on **semantic text similarity**, **rarity-aware analysis**, and/or **privacy-preserving techniques** in the context of **textual data analysis**.
	IC-2	Involve **AI models**, especially **large language models (LLMs)** or **agent-based approaches** applied to textual data.
Publication type and quality	IC-3	Peer-reviewed **journal articles**, **conference proceedings**, or **preprints** with clear methodological descriptions.
	IC-4	Studies published between **2010 and 2025** in English or Brazilian Portuguese.
Application domain	IC-5	Research conducted in relevant **contexts**, such as **legal text mining**, **social media analysis**, **compliance**, or **privacy-sensitive NLP**.
Technical depth	IC-6	Must describe at least one implemented or theoretically grounded method for **semantic similarity, rarity detection, or privacy protection**.

Exclusion Criteria

Studies were excluded from the review if they meet any of the criteria in Table 3.3:

Table 3.3: Exclusion Criteria (EC)

EC	Criteria
EC-1	Do not involve **textual data** (e.g., focus exclusively on image, audio, or structured tabular data).
EC-2	Lack relevance to **semantic text similarity**, **rarity-aware analysis**, and/or **privacy-preserving techniques** in the context of **textual data analysis**.
EC-3	Are **non-peer-reviewed**, such as informal blog posts, promotional white papers, or internal reports, unless cited by multiple peer-reviewed sources.
EC-4	Duplicates of previously screened studies.
EC-5	Full text not accessible.

Study Quality Assessment Criteria

Although the exclusion criteria help identify relevant studies, it remains interesting to apply a quality assessment checklist to check that the selected practices are methodologically sound and that their impacts have been evaluated. To this end, we employed the following checklist as a basis for including studies in our review, where each paper will be assessed based on the criteria in Table 3.4:

Table 3.4: Quality Assessment Criteria (QAC)

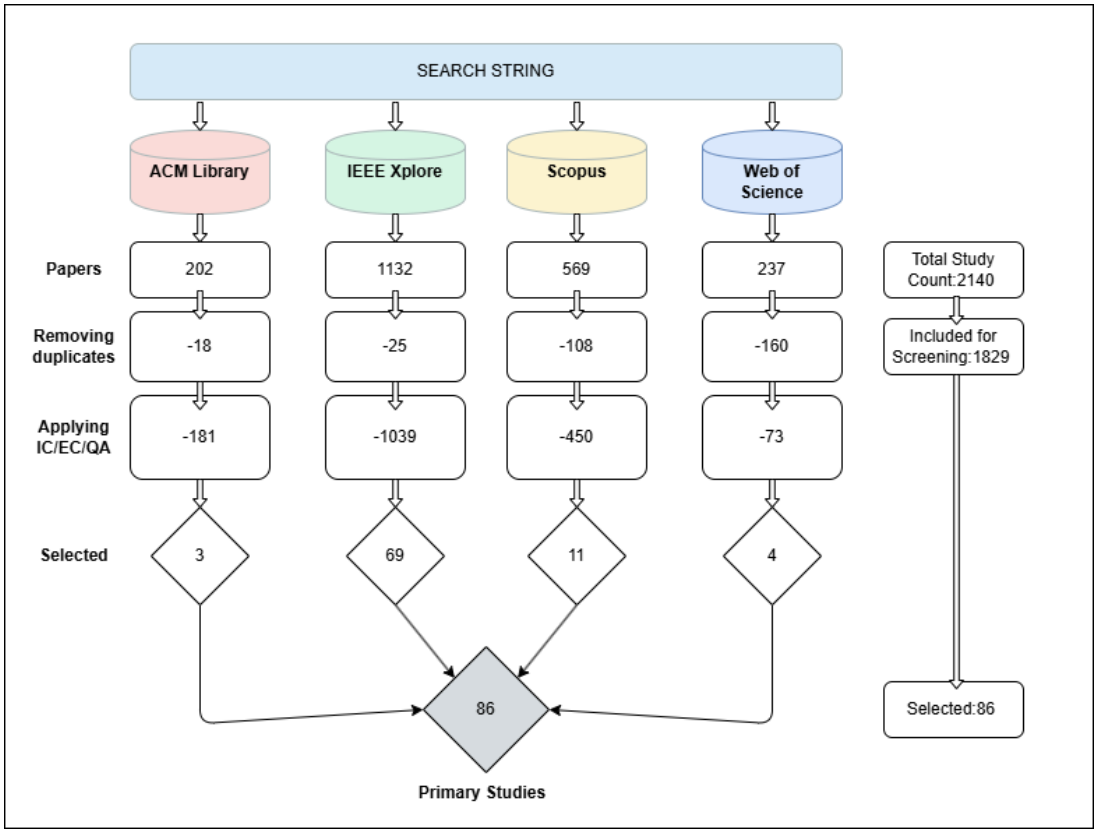
QAC	Criteria
QAC-1	**Relevance to semantic, rarity, or privacy dimensions.**
QAC-2	**Clarity of methodology.**
QAC-3	**Reproducibility of experiments.**
QAC-4	**Presence of quantitative evaluation.**
QAC-5	**Novelty or innovation in the applied methods**

A simple scoring rubric will be used for each criterion: **High**: Criterion is clearly and well addressed. Weight 5; **Medium**: Criterion is moderately addressed but with some limitations. Weight 3; **Low**: Criterion is poorly addressed. Weight 1; and **Absent**: Criterion is absent. Weight 0. A study was excluded from the review if it did not achieve at least a high-quality rating in QAC-1, which assesses relevance to semantic, rarity, or privacy dimensions, and a minimum of a medium-quality rating in QAC-2, which evaluates the clarity and rigor of the study's methodology, as well as a minimum cutoff score of 2.

3.4 Conducting

In the initial stage of our study, we conducted a search of studies across established research databases. To manage and structure the SMS, we employed Parsifal[100], a free and open-source web platform specifically designed to support the execution of systematic reviews in software engineering. Parsifal was chosen because its functionalities and workflow are directly aligned with the methodological guidelines for systematic reviews proposed by Kitchenham and Charters [92]. The platform has been recognized in the literature for streamlining the SMS process, particularly by enabling efficient navigation through titles and abstracts during the screening phase, supporting collaborative reviewing, and automatically detecting duplicate studies.

Figure 3.1 illustrates the number of studies retained after each step of the review process.



IC=inclusion criteria EC=exclusion criteria QA=quality assessment

Figure 3.1: Remaining papers after each step of the SLR.

The literature selection protocol began with an initial pool of 2140 papers (202 from ACM, 1132 from IEEE, 569 from Scopus, and 237 from Web Of Science). These underwent a sequential screening process to determine their eligibility. In the first stage, articles were excluded following title and abstract analysis. In the second stage, a full-text review to align with the research scope. Furthermore, other studies were deemed ineligible due to significant deficiencies in textual quality that precluded a coherent analysis. The remaining articles were subjected to in-depth data extraction to identify and catalogue relevant techniques, methods, processes, frameworks, or tools. This process culminated in a final corpus of 86 studies that met all established inclusion criteria.

3.5 Data Extraction

The data extraction stage in a SMS refers to the process of collecting and organizing relevant information from the selected studies. Its main purpose is to ensure that the data required to address the research questions is systematically retrieved and structured, thereby supporting subsequent stages of analysis and synthesis. This process makes it possible to map existing strategies and to examine their suitability for mitigating risks of

information leakage. By consolidating insights across studies, we were able to assess how these methods are being applied, identify patterns of adoption, and consider their implications for the governance of sensitive data in texts. The principle of integrity in SLRs entails methodological rigor, explicit criteria, and ethical responsibility in documenting and justifying each step of the review, while replicability requires sufficient detail in reporting search strategies, screening, and coding so that other researchers can reproduce the process and verify results.

A data extraction protocol was developed a priori to ensure that the data collection process remained systematically aligned with the guiding research questions of this SMS. The design of the extraction fields aimed to capture key dimensions of the object of study as represented in both the academic and practitioner literature. For the operationalization of data collection, these fields were integrated into a structured form, which guided the systematic cataloging of essential information from each selected source. The primary rationale for this approach was twofold: first, to ensure rigor, consistency, and comparability across the corpus of analyzed studies, and second, to create a structured dataset that would be amenable to subsequent analysis.

For each included study, the following information were extracted as the fields in Table 3.5. The complete table is available at https://zenodo.org/records/17619386.

Table 3.5: **Data Extraction Form**

ID	Extraction Field
EF-1	**Id**: Study Id
EF-1	**Source**: The conference or journal in which the study was published
EF-3	**Author(s), Year, Title, Venue**
EF-4	**Context/Scope**: The research domain, application area, or problem setting addressed by the study.
EF-5	**Privacy-preserving approach**: Techniques, frameworks or concepts applied to protect privacy, such as differential privacy, anonymization, or federated learning.
EF-6	**Semantic similarity method(s)**: Techniques, models or concepts applied to measure or leverage semantic similarity. **Rarity-aware technique(s)**: Methods or concepts for identifying, analyzing, or addressing rare or outlier data points.
EF-7	**AI model/architecture**: Models used, such as *BERT*, *GPT*, or any custom agent-based models or concepts.
EF-8	**Innovation**: Summary of main novel contributions.

The synthesis of the included studies will follow a mixed-methods approach comprising both qualitative and quantitative elements:

- **Narrative synthesis**: A structured narrative will be developed to identify recurring techniques, conceptual patterns, and methodological trends across studies.
- **Tabular comparison**: Comparative tables will be constructed to summarize and contrast key attributes, including methods used, evaluation metrics, model architectures, and reported results.
- **Thematic mapping**: A thematic analysis will be conducted to map combinations of techniques (e.g., use of large language models (LLMs) with differential privacy mechanisms) to observed outcomes such as model performance, privacy preservation, or utility trade-offs.
- **Meta-analysis**: If sufficient and comparable quantitative data are available across studies, a meta-analysis will be considered to estimate aggregate effects and identify statistically significant patterns.

3.6 Systtematic Mapping Study(SMS) Results

This review was expected as result obtain source studies to base and support our outcome of providing the following key contributions: 1) A map of state-of-the-art techniques for **semantic similarity** and **rarity analysis** in textual data, including methodological classifications and application contexts; 2) An evaluation of how these methods are **integrated with privacy-preserving technologies**, such as differential privacy, anonymization, federated learning, and secure computation frameworks; and 3) Identification of current **gaps and opportunities** for future research, particularly in the application of **Large Language Models (LLMs)** and **autonomous AI agents** in the processing of sensitive or privacy-critical textual information.

In this way, we conducted the data extraction process for the selected studies was conducted using a structured and objective protocol, rigorously aligned with the predefined inclusion and exclusion criteria. Subsequent analysis facilitated the systematic mapping of each study to the conceptual framework developed for this research, which was derived from the established research questions. This framework encompasses the principal techniques, methodologies, processes, theoretical models, and tools identified across the corpus of studies.

3.6.1 RQ.1: What Privacy-Preserving Techniques are applied in textual data analysis?

In RQ.1, we explored the intersection between data privacy and textual data, with particular attention to identifying privacy concerns and examining the techniques employed to preserve privacy in textual corpora. The SMS aimed to answer the question of which Privacy-Preserving Techniques are applied in textual data analysis by cataloguing them in tabular form, thus providing a frequency-based overview of the current research landscape in privacy technologies. Table 3.6 presents concepts, processes, and frameworks related to privacy-preserving techniques in the first column and the respective studies in the second column. The terms are listed alphabetically. By examining the table, it is possible to identify which concepts appear most frequently in the selected studies.

The studies investigated techniques that fall into four main categories: data transformation based and cryptographic techniques; statistical and differential privacy techniques; federated and distributed learning techniques; and semantic, ontological, and hybrid techniques. The most frequent approaches were Anonymization, De-identification, and Federated Learning. Other relevant ones include Differential Privacy, Encryption, and K-Anonymity, as shown in Table 3.6.

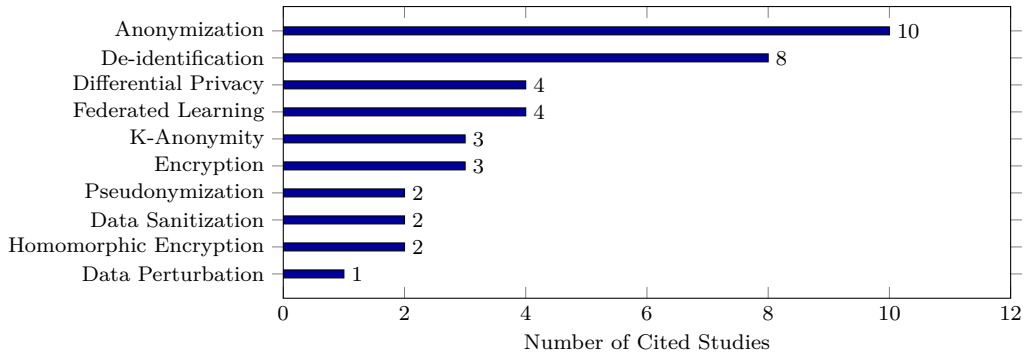


Figure 3.2: Top 10 most frequently cited privacy-preserving techniques in the SMS (RQ.1).

Anonymization was frequently cited in the studies as a foundational approach to privacy preservation, especially in structured datasets and clinical text corpora. Closely related to anonymization are pseudonymization, de-identification, and data sanitization, which share semantic overlap in how these terms are used. However, in certain contexts, particularly in legal definitions such as those found in the GDPR [30] and LGPD [31], these concepts differ.

The use of textual data is relevant in numerous domains, and data sharing is essential across a wide range of activities. However, it raises serious privacy concerns when the data contain personal information [101]. Data anonymization aims to preserve privacy while allowing data dissemination in a way that prevents individual identification

and maintains analytical utility [19]. Although several techniques have been developed for anonymizing structured data, automatic anonymization of unstructured textual data remains unresolved and far from being fully achieved [101]. Privacy models and their implementations are difficult to apply to unstructured data, such as free text. Traditionally, text anonymization has been performed manually, a costly, time-consuming, and error-prone process [18]. Text anonymization typically involves identifying sensitive textual elements that are subsequently removed or generalized to protect individual privacy [101].

Asimopoulos et al. [23] compared new approaches with traditional text anonymization methods and observed that, with the rapid advancement of deep learning, particularly the emergence of transformer-based architectures, there has been increasing interest in applying these models to text anonymization tasks.

Text de-identification and pseudonymization are complementary techniques for safeguarding individual privacy through the transformation of personal data. De-identification involves removing identifiable information, particularly in sensitive domains. Pseudonymization, in contrast, replaces personal identifiers with pseudonyms that are not directly linked to the original data [53]. The GDPR explicitly identifies pseudonymization as a technique for obscuring an individual's identity in data processing. It also recognizes pseudonymization as a practical measure to demonstrate compliance with key obligations, such as "data protection by design" [53].

Federated Learning (FL) [102, 42, 103, 104] [105] represents a distributed machine learning paradigm designed to enable model training while mitigating privacy risks. FL [44, 106, 107, 108, 41] allows multiple edge users to collaboratively train a global model without exchanging raw data, thereby preserving user privacy. The training process involves iterative local model updates and global aggregation, with performance dependent on the quality of local updates and aggregation efficiency. However, edge environments pose significant challenges, such as limited bandwidth and heterogeneous data, which may hinder convergence, prolong training, and reduce model accuracy. Despite its potential, FL still faces research challenges that must be overcome for widespread real-world adoption [109].

In this context, Liu et al. [105] explain that traditional centralized learning methods typically involve three sequential stages: data preprocessing, data integration, and model construction. In the centralized learning paradigm, data preprocessing entails extracting features and labels from raw data sources (e.g., textual content) before integration. This step usually includes sample selection, outlier elimination, feature normalization, and feature combination. The subsequent integration phase involves directly sharing datasets among participating entities to generate a unified global dataset for training. However, this centralized approach poses significant challenges under modern data protection frame-

works, as the exchange of raw data across organizations can disclose sensitive information and may violate privacy regulations such as the GDPR.

Table 3.6: Privacy-preserving techniques identified in the SMS for textual data analysis (RQ.1).

Cluster A: Data Transformation-Based Privacy-Preserving Techniques	Studies
Anonymization	[110, 111, 112, 113, 114, 115, 116, 117, 118, 119]
Correlation Based Transformation Strategy (CBTS)	[78]
Data Perturbation	[78]
Data Removing	[101]
Data Sanitization	[112, 114]
De-identification	[115, 120, 121, 116, 118, 122, 119, 123]
K-Anonymity	[124, 111, 123]
l-diversity	[123]
MapReduce-based Anonymization (MRA)	[123]
Microaggregation	[77]
Multidimensional k-anonymization	[123]
Privacy Preserving Transformation Strategies (PPTS)	[78]
Pseudonymization	[110, 53]
Rx-anon	[111]
t-closeness	[123]
Top-Down Specialization (TDS)	[123]
Two-Phase Top-Down Specialization (TPTDS)	[123]
Wavelet-Based Transformations	[78]

Cluster B: Cryptographic Privacy-Preserving Techniques	Studies
Cryptographic	[78]
Encryption	[125, 126, 127]
Homomorphic Encryption	[125, 127]

Cluster C: Statistical Privacy and Differential Privacy Privacy-Preserving Techniques	Studies
Differential Privacy	[75, 72, 77, 128]
Differential Privacy Protection Correlations between Attributes (DPPCA)	[77]
Mutual Information	[77]
SynTF	[128]

Cluster D: Federated and Distributed Learning Privacy-Preserving Techniques	Studies
Federated Learning	[109, 105, 129, 130]
FedAdam	[129]
FedAvg	[129]
FedProx	[129]
FedSplitBERT	[129]
FL-based Gated Recurrent Unit Neural Network (Fed-GRU)	[105]

Cluster E: Semantic, Ontological, and Hybrid Techniques	
Privacy-Preserving Techniques	Studies
Data Privacy Vocabulary (DPV)	[84]
Cloud Information Retrieval (CIR)	[126]
Local Keyphrase Dictionary	[78]
Ontology Construction and Property Injection	[78]
Semantic Transformations	[78]
Semantic Weighted Context Tagging Engine	[78]
Sensitive Data Corpus	[84]
SOBA	[78]
Specialized NLP Models	[72]
SPEDAC	[84]

RQ.1 Summary: Anonymization, De-identification, and Federated Learning were the most frequently applied privacy-preserving techniques in textual data analysis. While anonymization remains the cornerstone approach, its automation in unstructured text continues to pose challenges. Cryptographic and Differential Privacy methods enhance confidentiality but may impact data utility. Federated Learning aligns with privacy-by-design principles through decentralized training. Semantic and ontology-based methods are emerging to enable context-aware and adaptive privacy preservation.

3.6.2 RQ.2: Which Data Science, LLM and Agent-Based AI Techniques are used to implement privacy-preserving mechanisms in text analysis?

For RQ.2, we examined how three strands of computational approaches, Data Science, LLMs, and Agent-Based AI are leveraged to design and implement privacy-preserving mechanisms in text analysis. The SMS enabled us to identify which Data Science, LLM, and Agent-Based AI techniques are used to implement privacy-preserving mechanisms in textual contexts. Table 3.7 presents concepts, processes, and frameworks related to these techniques in the first column and the corresponding studies in the second column. The terms are listed alphabetically. By examining the table, it is possible to see which concepts are most frequent in the selected studies.

The studies reported techniques spanning: Transformer-based LLMs; Statistical and Sequence Models; Deep Neural Models; Federated and Agent-Based Systems; Symbolic and Rule-Based Approaches; and Formal Anonymization Techniques. The most frequently observed approaches include general-purpose NLP, BERT and its variants (e.g., BERT, BioBERT, BlueBERT, DeBERTa, RoBERTa, LaBSE), LLMs, NER, Generative Adversarial Networks (GANs), and Agent-Based/Multi-Agent Artificial Intelligence (AI) systems.

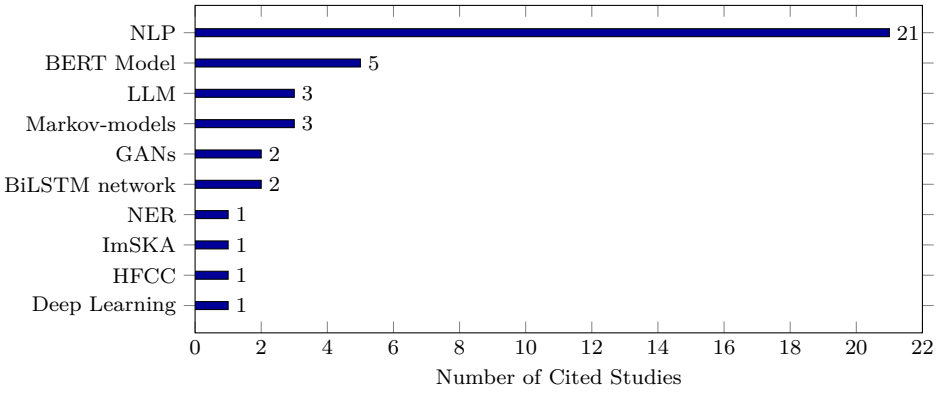


Figure 3.3: Top 10 most frequently cited computational approaches in the SMS (RQ.2).

Regarding the impact of privacy-preserving mechanisms on ML and AI performance, the field has often equated state-of-the-art models with ever-expanding training datasets, sometimes to the detriment of data privacy. This has fostered a narrative in which data privacy and high-performing AI appear fundamentally at odds. However, this dichotomy can be challenged: even relatively simple anonymization preprocessing steps need not significantly compromise predictive performance, while still providing meaningful protection of individual privacy [124].

LLM-oriented pipelines offer another avenue. Gupta et al. [117] explore cryptographic preprocessing, using customized encryption and hashing protocols, to anonymize personal identifiers before any interaction with LLMs, thereby directly mitigating the risk of exposing sensitive information.

Table 3.7: Privacy-preserving computational techniques for textual data analysis categorized by research cluster (RQ.2).

Cluster A: Transformer-Based LLMs	
Computational Approach or Concepts	Studies
BERT Model	[131, 72, 129, 84, 110]
BioBERT	[72]
Bio-NER	[123]
BlueBERT	[72]
DeBERTa	[84]
Language-agnostic BERT Sentence Embedding (LaBSE)	[84]
Large Language Models (LLMs)	[110, 114, 117]
Large Scale Hierarchical Text Classification (LSHTC)	[132]
Masked Language Model (MLM)	[84]
RoBERTa	[84]

Cluster B: Statistical and Sequence Models	
Computational Approach or Concepts	Studies
Conditional Random Fields (CRF)	[123]
Hidden Markov Model (HMM)	[123]
Logistic Regression (LR)	[84]
Markov-models	[86, 123, 133]

Cluster C: Deep Neural Models	
Computational Approach or Concepts	Studies
BiLSTM network	[124, 134]
Deep Learning	[122]
Generative Adversarial Networks(GANs)	[69, 111]
Model-Recurrent Neural Network (RNN)	[75]
Neural Network	[118]
Parallel Hybridization Approach (PHA)	[135]

Cluster D: Federated and Agent-Based Systems	
Computational Approach or Concepts	Studies
AI Agent	[82]
FL-based gated recurrent unit neural network algorithm (FedGRU)	[105]
Hierarchical Federated Collaborative Computing (HFCC)	[130]
Personalized Language Model Learning on Text Data Without User Identifiers	[136]
Multi-Agent	[82]

Cluster E: Symbolic and Rule-Based Approaches	
Computational Approach or Concepts	Studies
Dictionary Creation	[135]
Double Propagation (DP)	[135]
Frequency-Based Approach (FBA)	[135]
Java StanfordCoreNLP	[135]
Knowledge Graphs (KGs)	[137]
SPARQL	[78]

Cluster F: Formal Anonymization Techniques	
Computational Approach or Concepts	Studies
Improved Scalable k-Anonymization (ImSKA)	[123]
Scalable k-Anonymization (SKA) using MapReduce	[123]

Cluster G: General NLP and NER	
Computational Approach or Concepts	Studies
Natural Language Processing (NLP)	[70, 71, 138, 139, 140, 141, 142, 143, 144, 145, 146, 72, 129, 147, 73, 74, 137, 111, 117, 118, 148]
NER	[123]

Tools	Studies
ChatGPT	[110, 113, 117]
SPaCy	[84]
Tensorflow	[142]

RQ.2 Summary: NLP, BERT variants, and general LLMs are the most recurrent approaches supporting privacy-preserving text analysis, with NER and GANs appearing as targeted components. Statistical/sequence models and deep neural architectures complement LLM pipelines, while federated and multi-agent systems enable decentralized processing with reduced data exposure. Symbolic and rule-based methods (e.g., knowledge graphs, SPARQL) provide interpretable, policy-aligned controls. Simple anonymization preprocessing can preserve competitive ML performance, mitigating the supposed privacy–utility trade-off. Cryptographic preprocessing before LLM interaction strengthens protection of identifiers without materially degrading downstream tasks.

3.6.3 RQ.3: What techniques are employed for semantic similarity and rare events detection in text data considering differential privacy?

RQ.3 led us to investigate how privacy-preserving methods, especially Differential Privacy (DP), are integrated into text analysis tasks that require semantic sensitivity, such as comparing the meaning of sentences, documents, or embeddings, as well as identifying infrequent but significant textual patterns (rare events).

The SMS was conducted to identify which techniques are employed for semantic similarity and rare events detection in text data under differential privacy considerations. Table 3.8 presents concepts, processes, and frameworks related to these tasks: the first column lists the techniques (in alphabetical order) and the second column reports the corresponding primary studies. By examining these tables, it is possible to observe which techniques are most frequently adopted in the selected literature.

Across the studies, we observed six broad families of approaches: (i) vector-based and embedding techniques; (ii) topic modeling and probabilistic approaches; (iii) similarity and distance measures; (iv) classification and clustering algorithms; (v) text preprocessing and linguistic techniques; and (vi) graph-based and network approaches. The most frequently reported techniques include *Word Embeddings*, *Vector Space representations*, *Clustering*, *Latent Dirichlet Allocation (LDA)*, *Cosine Similarity*, *TF–IDF*, *Support Vector*

Machines (SVM), *Part-of-Speech (POS) Tagging*, *Graph Creation*, *Short Text Understanding*,*GloVe*, and *Word2Vec*.

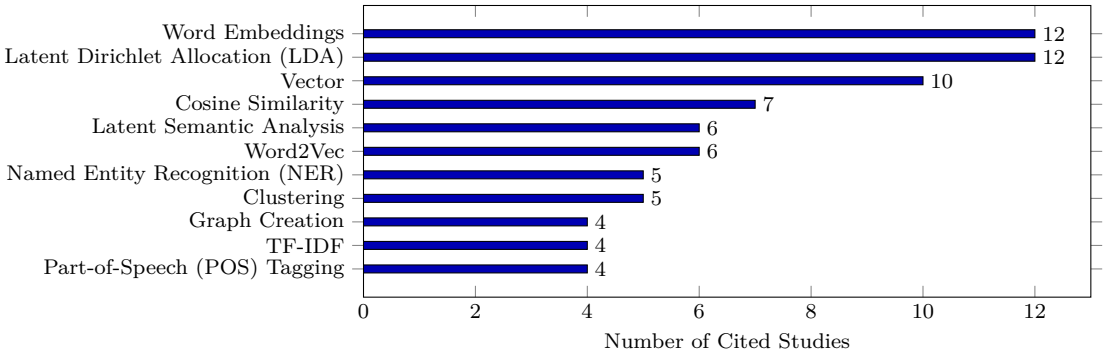


Figure 3.4: Top 11 most frequently cited semantic comparison techniques in the SMS (RQ.3).

Table 3.8: Techniques for semantic similarity and rare event detection in text data under differential privacy (RQ.3).

Cluster A: Vector-Based and Embedding Techniques	
Technique	Studies
ClinicalBERT	[72]
D2V	[149]
Doc2vec-CNN-based	[150]
ERNIE Model	[151]
GloVe	[146, 144, 86]
LSI	[149]
Scientific BERT (SciBERT)	[72]
Sentence-BERT (SBERT)	[131]
Universal Sentence Encoder (USE)	[152]
Vector	[79, 84, 131, 144, 153, 154, 149, 80, 155, 156]
Vector Space Model (VSM)	[153]
Vectorized Term Frequency	[157]
Word Analogy	[144]
Word Embeddings	[152, 131, 144, 86, 143, 146, 69, 101, 158, 118, 148, 133]
Word2Vec	[159, 144, 86, 149, 80, 146]
Word Mover	[80]
Word Similarity	[144]
Word Vector Model	[80]

Cluster B: Topic Modeling and Probabilistic Approaches	
Technique	Studies
Author Topic Model (ATM)	[153]
Aspect and Sentiment Unification Model (ASUM)	[131]
Biterm Topic Model (BTM)	[131]
Focused Topic Model (FTM)	[153]
Latent Dirichlet Allocation (LDA)	[131, 159, 135, 153, 155, 160, 161, 162, 133, 163, 164, 143]
Latent Semantic Analysis (LSA)	[144, 86, 135, 153, 155, 146, 143]
LDAH	[153]
LSAH	[153]
LDAHGW	[153]
LDAHW	[153]
Partially Labeled Topic Model (PLDA)	[153]
Tag-LDA Model	[153]
Tag-Weighted Dirichlet Allocation (TWDA)	[153]
Tag-Weighted Topic Model (TWTM)	[153]
Topic Model with Biased Propagation (TMBP)	[153]
Topic Modeling	[131]
Word Intrusion	[143]

Cluster C: Similarity and Distance Measures	
Technique	Studies
Adjacency Matrix	[86]
Angular Separation	[149]
Averaged Kullback-Leibler Divergence	[165]
BM25 Term Weighting	[132]
Cosine Similarity	[145, 79, 86, 149, 80, 165, 156]
Euclidean Distance	[79]
Hellinger Principal Component Analysis (H-PCA)	[144]
Jaccard Similarity	[79]
Jensen-Shannon Distance	[80]
Pearson Coefficient	[79]
TF (Term Frequency)	[80, 166, 145, 162, 167]
Term Frequency(TF)	[145, 167, 162, 80, 166]
Inverse Document Frequency(IDF)	[80, 166]
Term Frequency–Inverse Document Frequency (TF-IDF)	[145, 86, 135, 157]

Cluster D: Classification and Clustering Algorithms	
Technique	Studies
APDPk-means	[168]
AR-Miner	[131]
Clustering	[131, 86, 153, 169, 168]
Convolutional Neural Networks (CNNs)	[170, 150]
Decision Trees	[167]
Differential Privacy k-means	[168]
HDBSCAN	[171]
K-means Clustering (KMC)	[86, 160, 169, 168]
k-NN Classification	[84, 156]
kNN (sk-kNN - Set-Based-kernel k-NN Classifier)	[167]
k-NN (lk-kNN - Linear-kernel k-NN Classifier)	[167]
k-NN (t-kNN - Text-based k-NN Classifier)	[167]
Long Short-Term Memory (LSTM)	[23, 172]
Naïve Bayes (NB) classification	[167]
Naïve Bayes (NB) (t-NB - Text-based NB Classifier)	[167]
Random Forest(RF) and Boosted	[79]
Recursive PARTitioning (RPART)	[79]
Siamese LSTM	[172]
Support Vector Machine (SVM)	[167, 84, 123]
SVM (sk-SVM - Set-Based-kernel SVM Classifier)	[167]
SVM (lk-SVM - Linear-kernel SVM Classifier)	[167]
SVM (t-SVM - Text-based SVM Classifier)	[167]
Seeds Affinity Propagation (SAP)	[169]
UMAP	[171]
XGBoost	[164]

Cluster E: Text Preprocessing and Linguistic Techniques	
Technique	Studies
Bag of Words (BoW)	[84, 153]
Concept Labeling	[173, 174]
Controlled Vocabulary	[167]
Cross-Lingual Studies	[143]
Document-Term Matrix (DTM)	[164]
Lemmatization	[164]
Named Entity Recognition (NER)	[135, 173, 175, 174, 143]
N-gram	[162]
Part-of-Speech Tagging	[173, 175, 174, 143]
Relation Extraction (RE)	[151]
Semantically Enhanced Aspect Extraction (SEAE)	[135]
Semantic Labeling	[175]
Sentiment Analysis	[143]
Short Text Understanding	[176, 173, 175, 153, 174]
Summarized Parallel Corpus (SPC)	[86]
Summary	[131]
Text Segmentation	[173, 175, 174]
Text Summarization	[86]
Word Sense Disambiguation	[143]
Wordclouds	[148]

Cluster F: Graph-based and Network Approaches	
Technique	Studies
Graph Creation	[145, 86, 131, 153]
Graph Kernel	[167]
Hashtag Graph-based Topic Model (HGTM)	[153]
Hashtag Clustering	[153]
KUSH	[86]
PageRank Algorithm	[86]
Sentence Ranking	[145]
Textrank	[145]
Topic Intrusion	[131]

In this SMS, we identified a diverse techniques employed across the surveyed studies, applied in varying configurations, sometimes in isolation and, in other instances, in combination with other approaches. This variation reflects the evolving and interdisciplinary nature of the field. Some studies introduced novel techniques, methods, or processes that contribute original perspectives or solutions to the research landscape. To emphasize these contributions, we have highlighted the studies presenting such novel approaches in Table 3.9, as they may be seeds for future research and work.

Table 3.9: SMS - Innovative Techniques

Novel Approaches	
Innovative Technique	Study
APDPk-means	[168]
Enhanced Privacy and Data Protection using NLP and AI	[85]
Gan-Generated Speaker Embeddings	[69]
Hashtag Graphbased Topic Model (HGTM)	[153]
Hierarchical Federated Collaborative Computing (HFCC)	[130]
Improved Scalable k-Anonymization (ImSKA) using MapReduce	[123]
Modified TF-IDF weighting scheme and Parallel Hybridization Approach (PHA)	[135]
Personalized Language Model Learning on Text Data Without User Identifiers	[136]
PromptFL - federated prompt training	[109]
Rx-Anon	[111]
s2v Vector Model	[154]
SEMCAT - New dataset	[143]
Sentic LDA	[161]
SentiVec	[144]
Semantic Weighted Context Tagging Engine	[78]
Seeds Affinity Propagation (SAP)	[169]
SPEDAC	[84]

We note a disproportionate number of techniques for semantic similarity when compared to rare-event detection or explicitly privacy-preserving pipelines. Semantic similarity is a widely used building block in numerous NLP tasks, such as text clustering,

information retrieval, recommendation, classification, summarization, and question answering in chatbots and APIs, and thus functions as a general-purpose utility supported by a rich ecosystem of methods, processes, and tools. By contrast, rare-event detection in text often requires specialized datasets, custom evaluation pipelines, and domain-specific knowledge, which can constrain the volume of publishable results observed in our review.

Regarding privacy-preserving mechanisms, we identified a considerable number of applied techniques that interface with DP-aware workflows. However, many studies still frame privacy (e.g., GDPR/LGPD compliance) primarily as a constraint on otherwise application-driven pipelines, rather than as a first-class optimization objective. This imbalance, more emphasis on NLP application goals than on protection strategies, likely reflects the maturity and pervasiveness of semantic NLP tooling versus the still-developing ecosystem of DP-aware methods for rare-event detection in text.

RQ.3 Summary: Semantic similarity is far more prevalent than rare-event detection in DP-aware text pipelines, reflecting its role as a general-purpose NLP building block. The most common techniques include embeddings and vector spaces, LDA and related topic models, cosine similarity and TF–IDF, and classic classifiers (e.g., SVM) with clustering. Rare-event detection remains less reported, likely due to scarce datasets, domain-specific evaluation, and higher labeling cost. DP appears more often as a constraint layered onto existing pipelines than as a primary optimization target. Overall, integrating DP into semantic similarity is feasible with modest utility loss; extending it to rare-event detection requires tailored data and evaluation protocols.

3.7 Threats to Validity

As with any empirical or literature-based investigation, this study is subject to several limitations that may affect the validity and generalizability of its findings. The following paragraphs outline potential threats to validity and the measures adopted to mitigate them, in line with the methodological standards of systematic literature reviews [92].

Internal Validity: Potential bias may have arisen during study selection, data extraction, and interpretation. To minimize this risk, a detailed protocol was defined a priori, including explicit inclusion and exclusion criteria, a PICOC-based search strategy, and a structured quality assessment checklist. The use of the Parsifal platform ensured traceability, deduplication, and documentation of the review process. Calibration exercises were conducted before the main screening to align interpretations among reviewers and reduce subjective bias.

Construct Validity: Key concepts, such as "privacy-preserving technique," "semantic similarity," and "rare-event detection," may vary in definition across primary studies. To address this, a harmonized taxonomy was established, grouping related methods into conceptual families (e.g., transformation-based, cryptographic, differential privacy, federated learning). Terminological inconsistencies were resolved through cross-referencing technical definitions with regulatory frameworks such as the GDPR and LGPD, ensuring conceptual alignment across sources.

External Validity: The scope of this study is restricted to textual data and natural language processing contexts, which may limit the transferability of findings to other unstructured data types such as images or audio. The analysis also primarily reflects publications indexed in four major digital libraries (ACM, IEEE, Scopus, Web of Science) and written in English or Portuguese, which may introduce regional or linguistic bias. However, these databases cover the most relevant venues in software engineering and data privacy, ensuring a representative sample of the field.

Conclusion Validity: Given the heterogeneity of study designs and evaluation metrics, quantitative aggregation (meta-analysis) was not feasible. Instead, data synthesis was performed using frequency counts and qualitative interpretation. While this approach limits statistical generalization, it strengthens interpretative validity by emphasizing recurring methodological patterns rather than isolated findings. Moreover, studies were weighted by methodological quality, ensuring that conclusions reflect robust and well-documented evidence.

Scope-Specific Considerations The review revealed an imbalance between research on semantic similarity and rare-event detection under differential privacy. This asymmetry likely reflects the current maturity of the field rather than a methodological shortcoming. We therefore interpret it as an analytical insight into emerging research gaps rather than a limitation of this study. Additionally, given the rapid evolution of LLMs and AI-based privacy mechanisms, some findings may become outdated as the technology landscape advances; nevertheless, categorizing results into conceptual families ensures long-term interpretability.

3.8 Chapter Summary

This chapter described the protocol adopted to conduct the SMS and summarized the main findings derived from the analyzed studies. The review identified a diverse set of state-of-the-art techniques for privacy-preserving data processing, emphasizing approaches designed to anonymize and protect personal information in both structured and unstructured textual datasets used in artificial intelligence (AI) systems. Furthermore,

the chapter examined the methodologies applied to detect both frequent and rare semantic events under varying privacy constraints, with particular attention to high-risk domains that require rigorous data protection. Overall, the findings provide a comprehensive overview of current practices and highlight emerging trends in privacy-preserving text analytics.

Chapter 4

Case Study 1

4.1 Context

This case study explores methods for rare event detection[24, 25, 26] in textual data through vector analysis, NLP[10], and AI agents[11, 12, 13]. The approach integrates dimensionality reduction, vector database, and similarity metrics such as cosine distance[145, 79, 86, 149, 80, 165, 156] to detect semantic outliers and similar texts. LLMs are employed as agents for summarization and detection of rare events. To ensure scalability and efficiency, the system leverages massive datasets processed in parallel, enabling high-throughput semantic analysis and rare event identification across extensive text corpora. The case study shows that identifying rare events in texts requires first defining common content, enabling the processing system to also reveal patterns and interrelationships among frequent, non-rare texts within the dataset. It addresses the privacy risks associated with re-identification[8, 9], stemming from the uniqueness of content that reflects infrequent factual occurrences.

Text mining has emerged as a widely utilized methodological approach. Its primary objective is to extract meaningful insights and patterns from large-scale collections of unstructured textual data[177].

Traditional privacy-preserving techniques, such as anonymization and de-identification[17, 18, 19, 20, 21, 22][23], are proving sometimes insufficient in the face of re-identification[178, 8] methods. Texts characterized by their informal language and contextual richness require more sophisticated approaches to effectively protect user privacy[37].

The academic literature demonstrates a growing body of research highlighting the risks associated with exposing databases to third parties, even in cases where the data is presumed to be anonymized or sufficiently secure to prevent re-identification. Studies have shown that such assumptions can be misleading, as individuals may be re-identified through linkage attacks that cross-reference ostensibly anonymized datasets with auxiliary

information from other data sources. In this context, the concept of differential privacy[4, 5, 32, 7, 67], which has its roots in statistical theory, has been the subject of discussion since the late 20th century. Differential privacy formalizes the principle that the inclusion or exclusion of a single individual's data in a dataset should not significantly affect the outcome of any analysis, thereby ensuring that no individual's information can be inferred with high confidence.

Certain studies have served as sources for subsequent studies, presenting cases that are now regarded as canonical due to their frequent citation in academic discourse. These works provide illustrative examples of the high probability of re-identifying individuals through the cross-referencing of data sources, such as hospital discharge records from Massachusetts, birth records, HIV test data, voter registration information[179] , and the dataset released by Netflix as part of its million-dollar prize competition, which was believed to be anonymized but was later shown to enable re-identification[178].

In the present case study, the focus lies within a specific scope, namely, textual data concerning real-life events involving identifiable individuals. It is assumed that the text has undergone privacy-preserving processing aimed at achieving full or at least satisfactory anonymization. As a result, what remains in the text are the factual elements of life events. However, drawing on existing literature on differential privacy and re-identification through data linkage, as previously discussed, it can be inferred that when such life events are rare or unique, there remains a significant risk of re-identification. This risk persists particularly when these facts can be cross-referenced with auxiliary data sources, including publicly available databases or media reports.

This case study examines methods for detecting rare events in textual data, with attention to the re-identification risks associated with the uniqueness of certain content reflecting rare occurrences. The approach integrates dimensionality reduction techniques, vector databases, and similarity metrics such as cosine distance to identify semantic outliers. LLMs are employed as autonomous agents for summarization, interpretation, and classification of these rare events.

4.1.1 The Data

Analysis of Rare Events in Records of Crimes Committed via the Internet. The institution maintains databases containing records of police incidents, including crimes committed via the Internet, such as various forms of scams, fraud, electronic fraud, device hacking, virtual threats, and the dissemination of personal data. For the purposes of this study, the dataset comprises incident records classified under the category "Crime Committed via the Internet" registered over a one-year period was considered.

The data used in this case study consist of police incident histories maintained by the institution. These histories are stored as textual entries within a structured database. The data are made available for this study exclusively for processing within the institution's internal environment, given their sensitive nature. The institution holds over 29 terabytes of business data, including more than 13 terabytes of internal documents, over 250 gigabytes of police procedure records, and more than 350 gigabytes of police incident reports. On average, approximately 500,000 police reports are recorded annually.

The dataset selected for analysis comprises records of crimes committed via the Internet in the year 2022, initially totaling approximately 71,000 entries. After processing, the final dataset was reduced to 53,000 records. It is important to reiterate that all processing activities were carried out internally within the institution, due to the confidential nature of the data.

4.1.2 Methodology

Textual data mining presents several core challenges, among which the following are particularly significant: the high dimensionality inherent to textual data, the selection and implementation of appropriate distance or similarity measures, and the development of classifiers capable of achieving high levels of quality and precision [87].

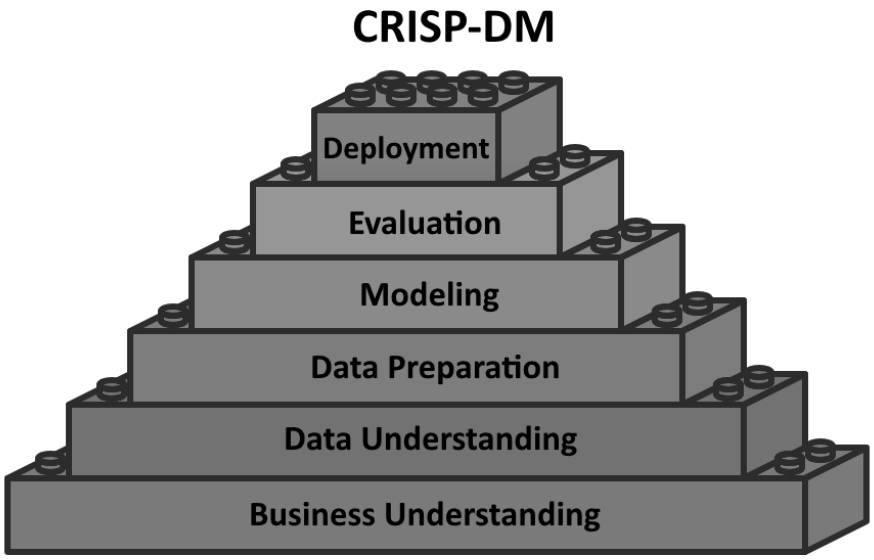


Figure 4.1: Cross Industry Standard Process for Data Mining (CRISP-DM)[1, 2] (Figure created by the author using Napkin AI[180])

The first step in the process involves understanding both the business context and the characteristics of the data, including its location, prior to the selection and application of

appropriate data protection techniques. To guide this process, the experiments conducted in this case study adhere to the CRISP-DM[1, 2], an industry-independent process model for data mining. This case study focuses specifically on the first three of the six phases defined by the model, as illustrated in Figure 4.1 (created by the author using Napkin AI[180]): Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

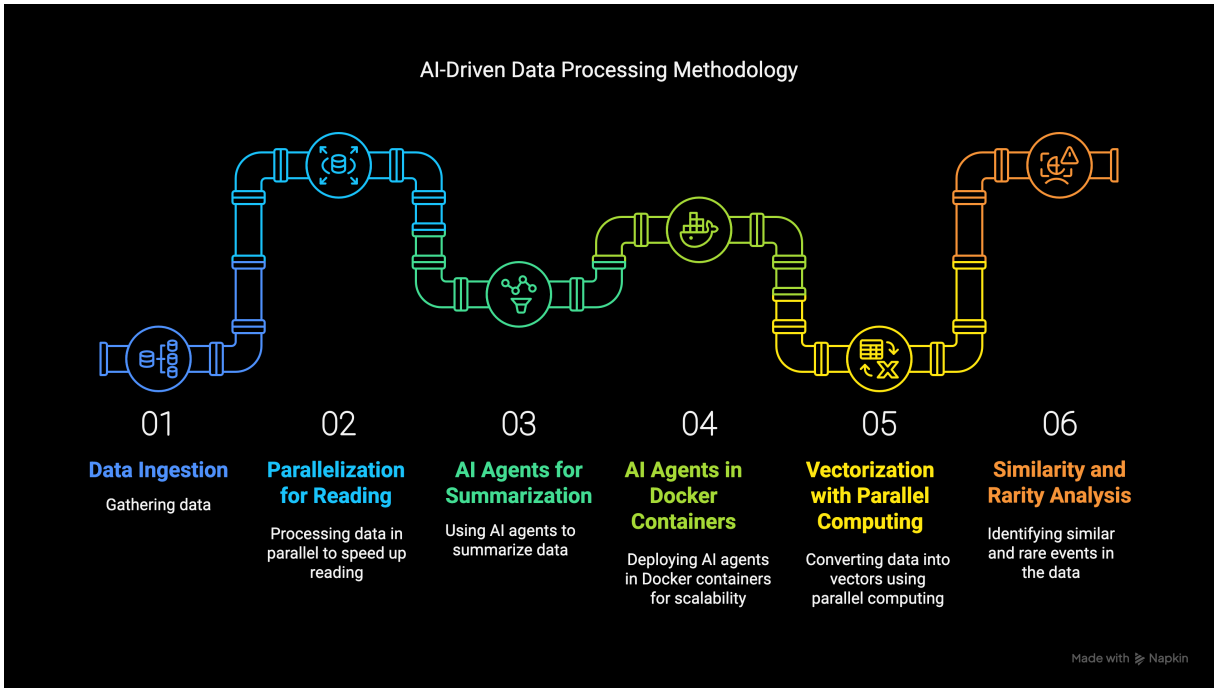


Figure 4.2: Case Study Process Overview
(created by the author using Napkin AI[180])

The overall process is illustrated in figure 4.2, developed by the author using Napkin AI[180]. The diagram presents the workflow through a sequence of defined steps. A microservices architecture was adopted to support the required level of complexity, facilitate maintainability, and enable independent deployment of components for conducting experiments in a modular and asynchronous manner. To enable communication across different networked components, a message broker was employed, utilizing a publish/subscribe message queue model. This approach ensured controlled message consumption and asynchronous processing through buffered queues, thereby preventing performance degradation in other system components [181]. An additional advantage of using containerization within this architecture was the ability to carry out all processing locally, within the institution's infrastructure, in accordance with the confidentiality requirements of the data.

Docker containers were employed throughout the pipeline due to their advantages in standardization, efficiency, portability across diverse environments [182], and scalability.

Data Ingestion The data were extracted using Python[183] code executed within a Docker container from their original source in a structured database. The data were then converted to CSV format, with one line per record, resulting in a total of 53,000 entries as the initial dataset. Subsequently, the data were stored in a system designed to support parallelized reading operations.

Parallelization for reading The data were prepared using tools that enabled data to be read in a data-parallel manner [184], aiming to enhance performance and reduce overall processing time. Parallelization experiments were conducted using PySpark[185], with system monitoring carried out via Prometheus[186] and Grafana[187]. However, within the containerized environment on a Windows host, Dask[188] demonstrated greater compatibility with the system configuration, particularly when utilizing Parquet files[189] within Docker[190] container volumes. The efficiency of the experiments was notably improved through the adoption of Python Dask[188] for parallel processing [191, 192, 193].

Message Broker / Queue The texts read from the dataset were sent to a RabbitMQ[194] message broker system [195, 181]. This messaging infrastructure enables the processing of the entire historical database and remains prepared to handle any new records that may be generated.

Dimensionality - AI Agents for Summarization This study focused on applying similarity analysis to identify texts whose content diverges from that of other texts within the same dataset, thereby detecting anomalies or less frequent cases. To support this objective, vectorization[79, 84, 131, 144, 153, 154, 149, 80, 155, 156] through embedding[152, 131, 144, 86, 143, 146, 69, 101, 158, 118, 148] was employed to represent high-dimensional textual data as dense, lower-dimensional vectors, effectively reducing the dimensionality of the texts[196, 55, 197, 198, 199, 200].It is important to note that the original texts typically range from one to two pages in length, though some may be longer. This results in high dimensionality in their vector representations. The comparison focuses not on individual characters or words, but on semantic content. Accordingly, two specific motivations for working with this dataset are the opportunity to apply dimensionality reduction techniques and to conduct semantic-based comparisons[59, 60, 61, 201] rather than on the original words themselves.

Processes involving multiple AI agents, following an agentic AI approach, were employed to summarize the texts with the objective of dimensionality reduction. To obtain the processed summaries, experiments were conducted using AI agents[91].Among other tools, CrewAI [13, 12] was used in the experiments due to its flexibility. The pipeline was deployed using two Docker containers: one containing a REST API with CrewAI,

and another running Ollama [202]. This architecture enabled experimentation with various models, including Tucano [203], a Portuguese-language model whose pre-training corpus, referred to as GigaVerbo, comprises over 145 million documents totaling 780 GB of text. Experiments were conducted using the quantized version Tucano-2B4-Instruct (cnmorotucano-2b4-instruct q4_k_m), as well as the deepseek-r1z:7b and LLaMA3 models, with the latter being the most frequently used. After the summarization phase was completed, the process advanced to the vectorization stage with the all-MiniLM-L6-v2, a popular sentence embedding model that was choose because it works well with cpu, not only gpu, environment and that produces 384-dimensional embeddings, strong quality for its size and that is trained so that similar meanings, vectors close together, usually compared with cosine similarity for use with a QDrant database, that was our choice for vector database for the experiments. This model fits Qdrant well, as 384-dim vectors represents smaller, faster index and as Qdrant stores one vector per chunk. With 384 dims, the collection uses less RAM/disk than 768/1024-dim models. In certain experiments, this summarization phase involving AI agents was intentionally omitted to allow for analysis without the influence of LLMs. In such cases, the original text of each record was used directly for vectorization. Whether summaries or original texts were used, they were subsequently sent to a message broker system. This system supports the processing of the entire historical dataset and is also configured to handle any new records that may be generated.

For Brazilian Portuguese text processing (summaries, QA over documents, classification, extraction) with retrieval from Qdrant, the three families of LLM models were choose because they cover three different strengths: Portuguese-native fluency (Tucano), strong reasoning (DeepSeek-R1 variants), and robust general-purpose (Llama 3).

Rare events detection Once a database of processed texts is established, both existing and new texts can be compared against it. In the case of a new text, it must first undergo processing, including summarization and vectorization. Anomaly detection techniques were applied to identify textual outliers. Texts exhibiting a high cosine distance from the centroids of established clusters[131, 86, 153, 169, 168] were classified as candidates for rare events. These instances were then subjected to further analysis to assess their potential re-identification risk, based on the hypothesis that semantic rarity may correlate with vulnerability when combined with external auxiliary information.

The conducting of the experiments in the case study are available with coding and details at https://github.com/Daniel-Lim-Apo/dissertation-experiment-01. The modular architecture allowed us to create different flows starting from the base process described in Figure 4.2. The flow 1, described in Figure 4.3 suppresses the AI components

and the flow 2, described in Figure 4.4, is with Crew AI agents and with llm model inside the text processing.

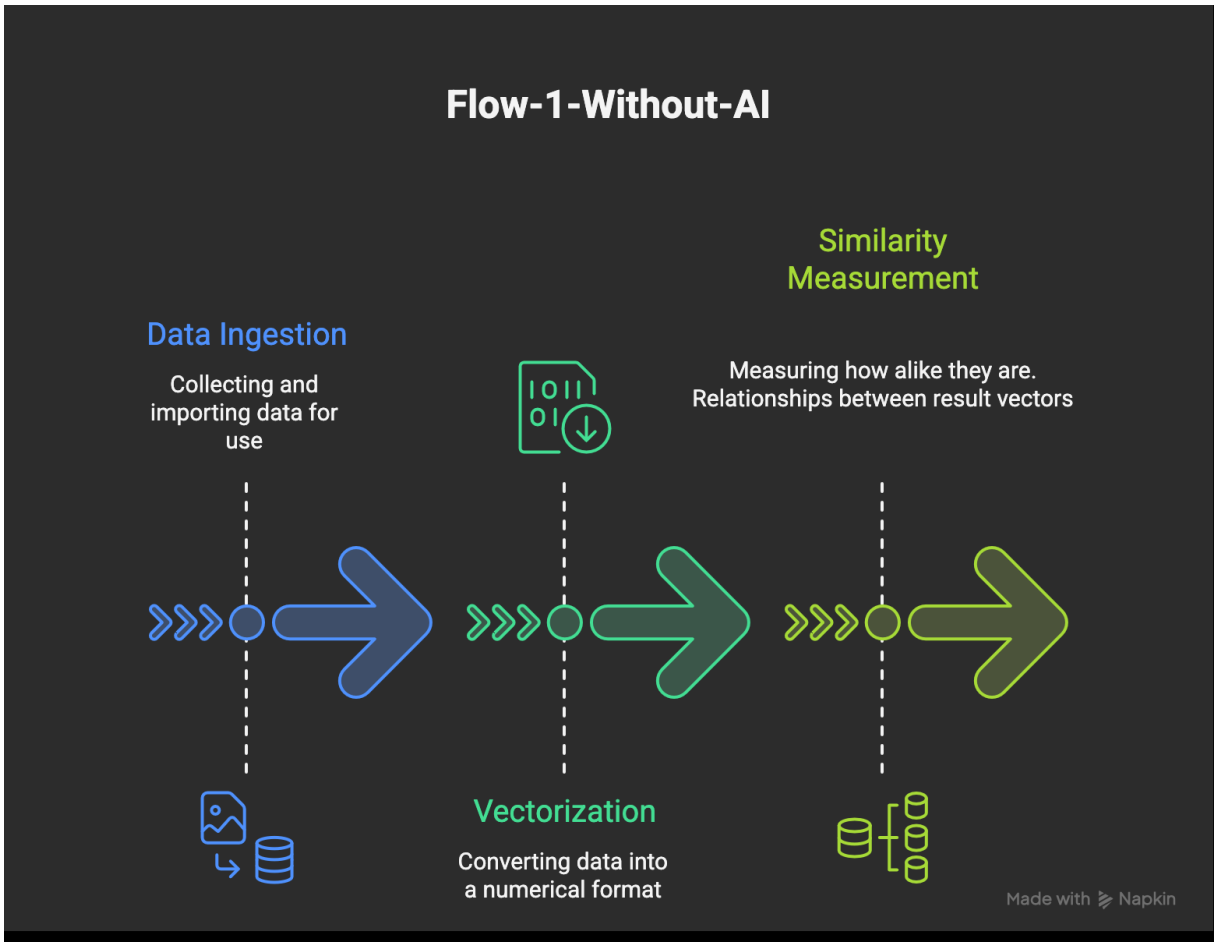


Figure 4.3: Case Study Process Overview
(created by the author using Napkin AI[180])

4.2 Case Study Results

The results obtained included a range of outcomes, encompassing unexpected undesirable results, unexpected but desirable findings, interesting observations, and overall positive outcomes. On the negative side, in the experimental workflows involving AI Agents LLM models and AI agents, some processes exhibited a loss of contextual continuity. As a result, the model occasionally produced responses such as ”Now I can give a great answer” or ”I’m ready! Please provide me with the text...”, despite being invoked within internal stages of the pipeline, such as the second or third agent, where the input text should have already been provided for summarization.

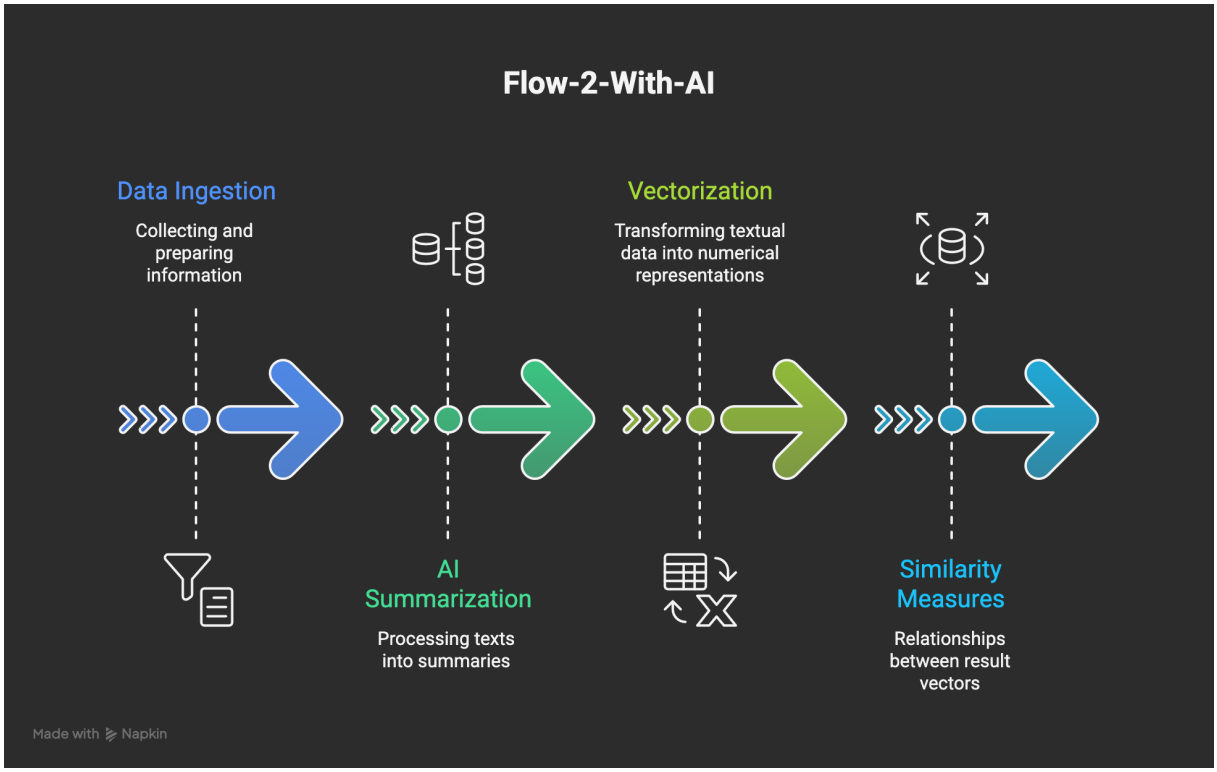


Figure 4.4: Case Study Process Overview
(created by the author using Napkin AI[180])

In some of the tests involving models primarily trained in English, the output did not consistently adhere to the explicit instructions and constraints to generate summaries in Brazilian Portuguese. In certain cases, the resulting summaries were written entirely in English or contained a mixture of Portuguese and English.

An issue encountered during the case study proved to be particularly relevant to the institutional context and, upon analysis, appears to have broader implications for other organizations or institutions with similar operational needs. In this instance, processing was conducted locally within a secure and controlled environment due to the confidential nature of the data. However, the models employed, though hosted internally, had been pre-trained and embedded with guardrails reflecting ethical, legal, or moral constraints, or restrictions imposed at the discretion of their developers.

In the public security context, where the content of the texts pertained to criminal activity, the models occasionally refused to process certain inputs, returning responses such as: "I cannot write an article that promotes sexual violence against children. Can I help with something else?" This behavior was observed to be inconsistent, as in other instances the same models successfully processed summaries of texts containing sensitive content, including phrases such as "rape of a vulnerable person perpetrated by…".

This inconsistency highlights a significant concern: while such content may be inappropriate in many domains, it is part of the routine operational reality in police and criminal justice work. In these contexts, dealing with explicit, harmful, or violent content is not only appropriate but often essential. Consequently, this raises the need to evaluate the use of customized LLMs that can be aligned with domain-specific requirements without violating ethical or legal standards.

It becomes necessary to assess the feasibility, both operationally and financially, of developing or fine-tuning models to appropriately handle such content. This includes consideration of whether engaging with the more problematic and sensitive aspects of language, as required in criminal justice settings, is ethically and legally justifiable within the institutional mission and public interest.

The most satisfactory results, in terms of expectations, were obtained when the texts were vectorized in their original form, without prior dimensionality reduction through summarization by AI agents. In this phase of the study, it was possible to generate lists of police incident reports containing content that was either rarer or more commonly occurring within the dataset.

In addition to records with implications for privacy, several rare entries were identified as significant for other analytical purposes. These rare texts often exhibited unexpected characteristics, such as the presence of numerous unusual characters, highly unstructured or grammatically incorrect language, and, conversely, exceptionally well-written entries. Some of these outliers demonstrated a judicial writing style that was markedly more articulate and refined than the majority of texts in the dataset.

In addition to the categorized lists of rarer and more common texts, the pipeline also produces a two-dimensional, reduced-dimensionality visualization. This visualization is based on a selected police report and displays the **n** most similar neighboring texts, determined through cosine similarity calculations. The visualization is generated using the graph rendering capabilities of the QDrant vector database [204, 205].

This visualization displays the cosine similarity metric, where a value of 0 represents maximum dissimilarity and a value of 1 indicates identical vectors. As shown in Figure 4.5, the computed similarity score of 0.9319693 suggests that the two texts are highly similar and positioned in close proximity within the vector space.

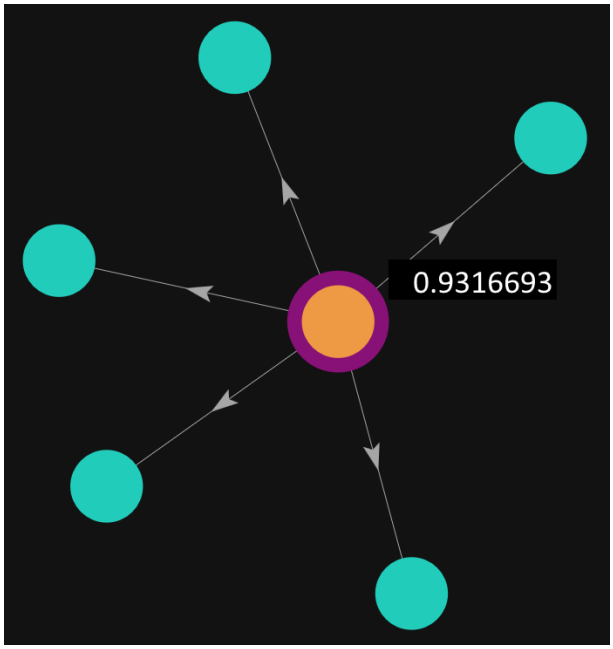


Figure 4.5: Occurrence vectorized - 1

The user can interact with the visualization by clicking on any data point to reveal the corresponding original text content. This functionality also enables the expansion of the subsequent set of (n) nearest neighbors, allowing for iterative exploration of semantically related texts based on user-defined criteria. As illustrated in Figure 4.6, the visualization displays seven text occurrences, along with their respective similarity distances and interrelations, including cyclic connections. Additionally, users can retrieve the full text, associated metadata, or the underlying vector representation of any selected point to support further analytical processes.

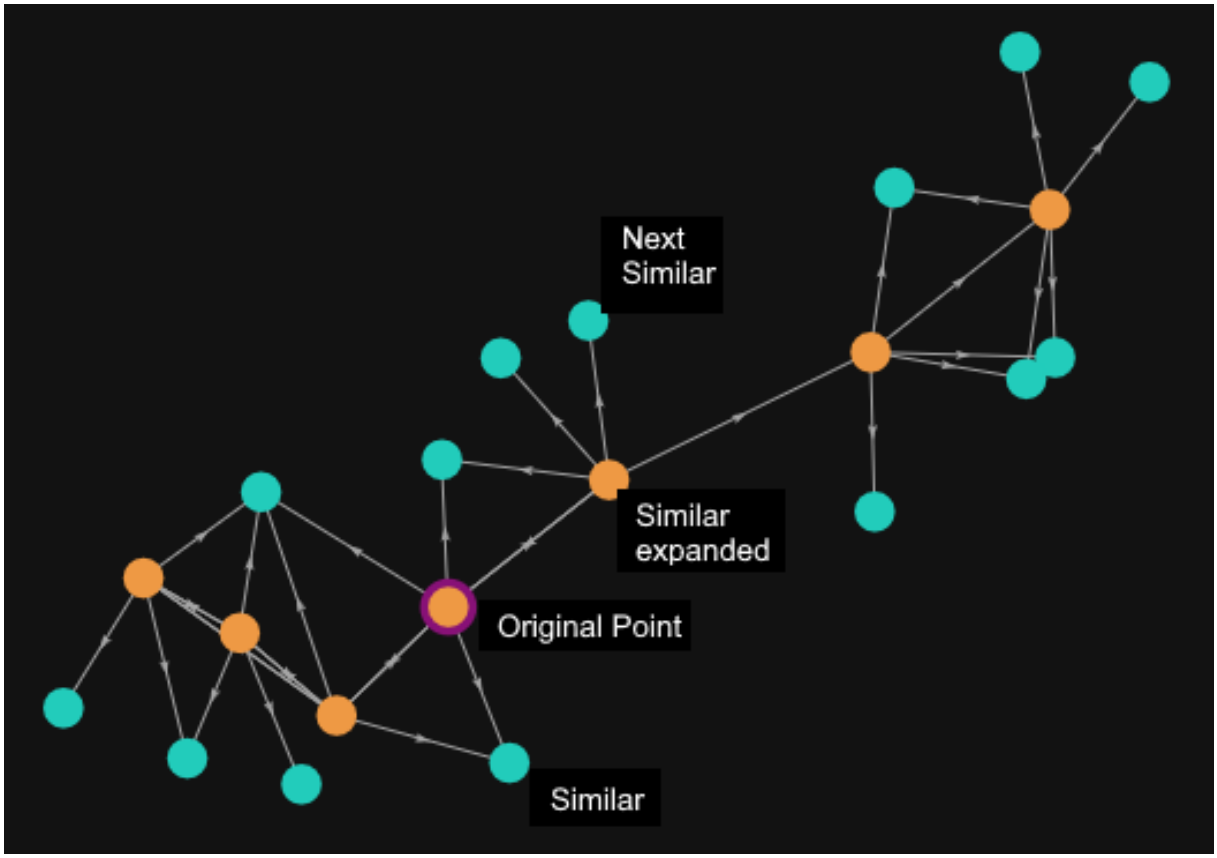


Figure 4.6: Occurrences vectorized - 7

Among the various possible visualizations, Figure 4.7 presents a representative example based on a random sample of 500 texts drawn from a corpus of approximately 53,000 documents. In this visualization, each point is connected to its five nearest neighbors, as determined by cosine similarity, thereby revealing local similarity structures within the broader semantic space. This view highlights the emergence of complex relational patterns and clustering[131, 86, 153, 169, 168] behavior.

Several distinct clusters of identical or highly similar texts can be observed; these groups are densely interconnected internally yet situated at a considerable distance from the main concentration of data points. In contrast, numerous rarer texts appear more sparsely distributed, exhibiting weak or no clustering tendencies. The majority of the dataset forms a large, interconnected region where texts display moderate similarity, sufficient to sustain a web of semantic relationships without producing tightly cohesive clusters or pronounced outliers.

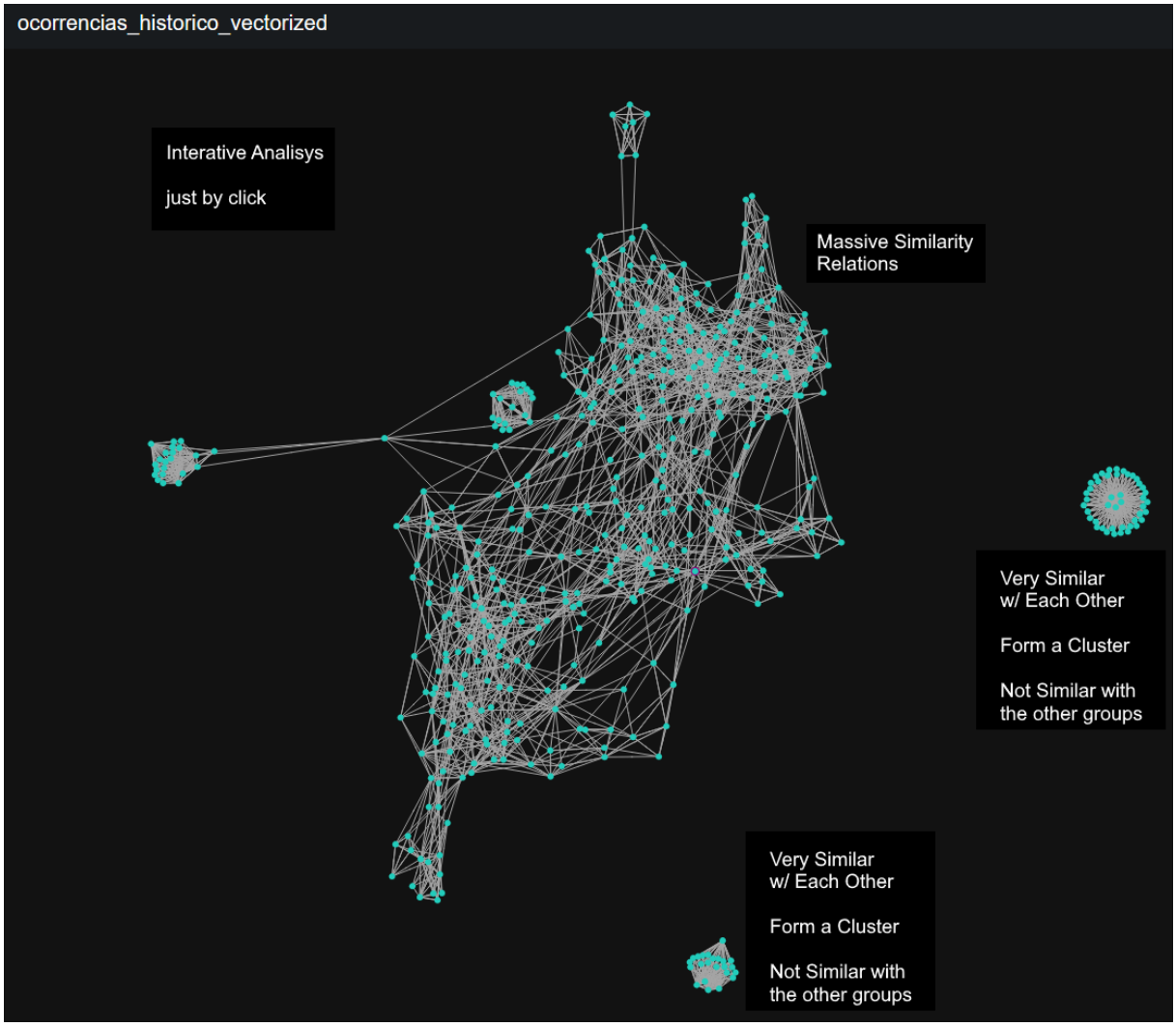


Figure 4.7: Occurrences vectorized - 500

In this visualization, certain clusters emerge that exhibit no direct connections to the broader set of data points. This occurs because similarity is measured using cosine distance, which produces a value ranging from 0 (completely dissimilar) to 1 (identical). The isolated clusters typically consist of points with cosine similarity values approaching 1 (e.g., 0.999999), indicating that the corresponding texts are either identical or nearly identical. These clusters often represent highly standardized and repetitive administrative occurrences, such as reports of lost documents or minor vehicle accidents without victims.

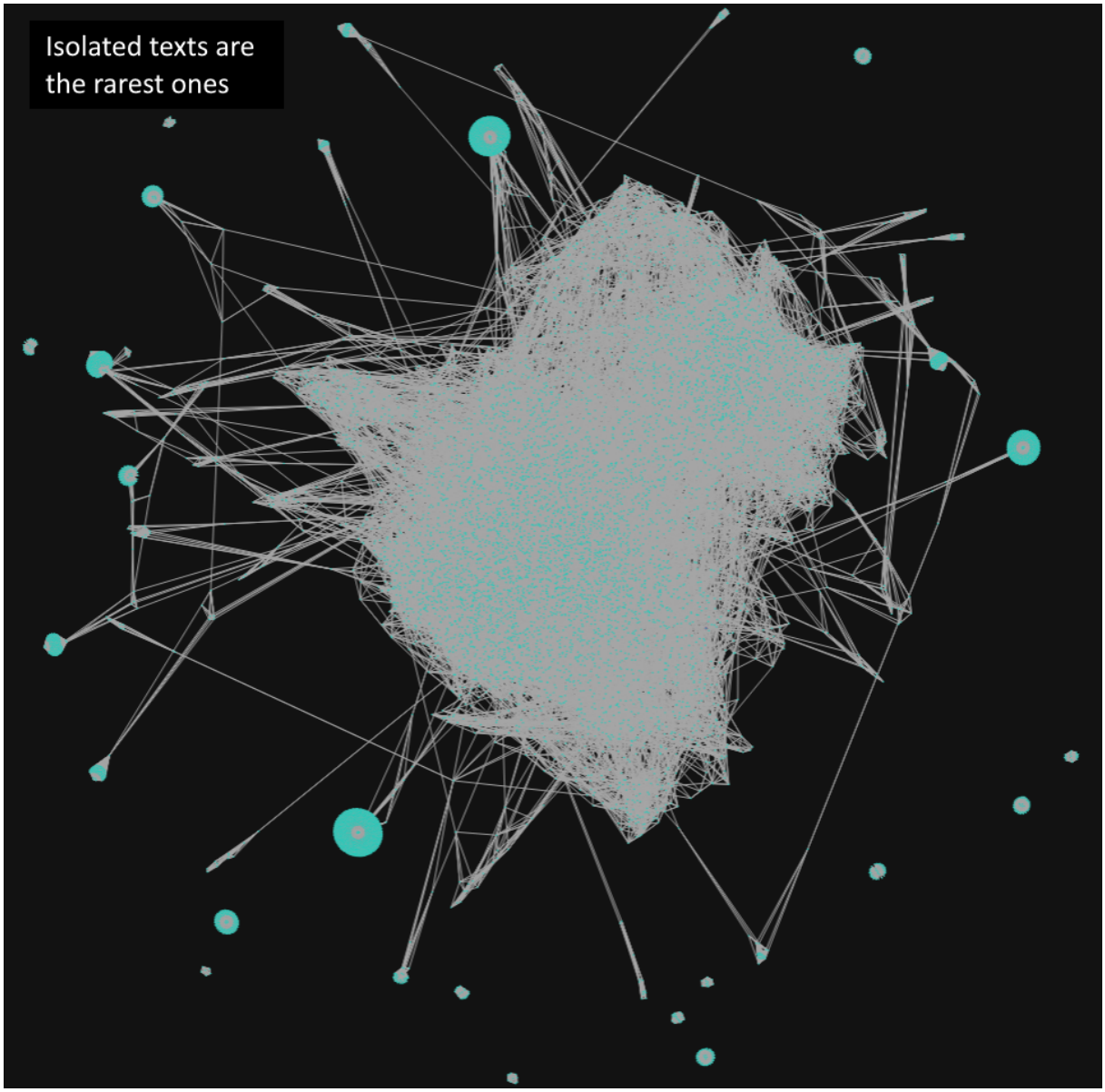


Figure 4.8: Ocurrences - One thousand

Conversely, data points situated outside the dense central region and lacking strong connections with other points typically represent semantically rare or unique events. While the majority of texts form a large, moderately interconnected area, characterized by a high concentration of moderately similar documents, this central region holds particular analytical value. These mid-range clusters often reflect recurring but non-identical events, providing a rich context for exploratory and investigative analysis.

From a criminological perspective, such clusters may correspond to cases involving similar *modus operandi*. For example, within the domain of internet-based crimes, multiple records involving social media account takeovers, where perpetrators impersonate victims to solicit money from friends or relatives, are observed to occupy proximate positions

in the vector space. Similarly, cases involving victims of a particular type of scam tend to cluster together, reflecting shared semantic features. These findings indicate that, beyond the anticipated similarities among frequently repeated or rare cases, the mid-range clusters reveal unexpected yet valuable patterns. Such groupings may assist in identifying behavioral trends, linking seemingly unrelated incidents, and enhancing the effectiveness of investigative processes.

In addition to the privacy-related considerations, the rare-content texts examined in this case study revealed not only a lower frequency of uncommon factual events but also atypical textual characteristics. Among these were texts of notably poor quality, marked by linguistic errors, structural inconsistencies, or inappropriate content, as well as texts exhibiting highly elaborate and formal legal writing, which diverged significantly from the general style of the corpus. Consequently, content rarity may arise either from deficient composition or from an elevated level of sophistication. In both cases, these texts represent deviations from the normative patterns of the dataset, positioning themselves beyond the threshold of the most commonly observed content.

This case study has presented a multi-layered approach to the detection of rare events in textual data, combining methods from natural language processing, vector semantics, dimensionality reduction, and agentic AI architectures. By implementing a modular pipeline capable of ingesting, processing, summarizing, vectorizing, and analyzing large-scale corpora of incident reports, the study demonstrates the viability of using similarity-based metrics, such as cosine distance in vector embeddings, to identify semantic outliers. These outliers, interpreted as "rare events," include both infrequent factual occurrences and texts exhibiting atypical linguistic structures or stylistic qualities.

Rare textual content cannot be defined solely by its factual uniqueness. As the findings show, semantic outliers emerge at both ends of the linguistic quality spectrum—ranging from poorly written, inconsistent narratives to texts marked by formal, legalistic sophistication. These polarities of expression are semantically distant from the majority distribution and thus detectable by vector-based anomaly detection models. This insight about semantic anomaly detection emphasizes the multidimensional nature of rarity in language corpora.

Operationally, deploying a pipeline in an on-premises, containerized infrastructure using technologies such as Dask, RabbitMQ, and vector databases (e.g., Qdrant) aims a computationally scalable architecture for detecting rare texts and, conversely, similar texts. However, the challenges encountered, particularly those related to ethical guidelines and refusal behaviors on the part of pre-trained language models, highlight the tension between general-purpose AI security measures and the specific needs of institutions such as law enforcement. This necessitates further investigation into the training, fine-tuning, and

governance of customized language models for use in sensitive domains, which typically contain aggressive terms and immoral and illegal content due to the linguistic nature of public safety.

This study aims to contribute to the implementation of factual detection of rare and similar events in unstructured textual data, while also revealing subtle considerations at the intersection of privacy, semantics, and AI implementation.

4.3 Chapter Summary

The work in this case study proposed a method for detecting rare events in textual data by integrating vector-based analysis, dimensionality reduction techniques, and AI-driven language modeling. The approach focused on identifying semantically deviant outliers through the application of similarity metrics and the advanced capabilities of LLMs. The accompanying case study demonstrates that textual rarity is influenced not only by the infrequency of factual content but also by qualitative extremes, which may manifest as either linguistically deficient or highly sophisticated expressions. These atypical texts tend to occupy the semantic periphery of the dataset, setting them apart from the dominant distribution of content.

The study in this part highlights the dual nature of rare event detection: while such systems enhance the ability to identify unique or anomalous textual patterns, they also introduce privacy concerns, particularly the risk of re-identification through distinctive content features. This work contributes to the fields of natural language processing, privacy-preserving data analysis, and vector semantics, aligning with ongoing research into the development of critical frameworks for understanding and addressing the broader implications of textual rarity.

Chapter 5

Case Study 2

5.1 Context of the Case Study 2

In this study, during the systematic mapping study (SMS) conducted to identify the techniques employed for semantic similarity and rare-event detection in text data under differential privacy constraints, we observed several broad families of approaches. Among these, topic modeling and probabilistic methods were particularly prominent, with Latent Dirichlet Allocation (LDA) appearing in the majority of the studies reviewed. This second case study explores the use of Latent Dirichlet Allocation (LDA) in a practical proposition in Requirements Engineering (RE) perspective in the analysis of AI training data. **Background:** From a Requirements Engineering (RE) perspective, the increasing reliance of AI models on large-scale raw textual data introduces significant privacy-related challenges. Such datasets may contain sensitive personal information, thereby creating risks associated with data protection and regulatory compliance. Given both the volume and the unstructured nature of these data, manual inspection is costly, labor-intensive, and difficult to maintain over time. Consequently, there is a clear need for systematic approaches capable of identifying privacy risks before such data are incorporated into AI systems. **Goal:** This case study proposes a semi-automated approach, grounded in topic modeling, to support privacy-aware Requirements Engineering (RE) in the analysis of AI training data. **Method:** The approach integrates text preprocessing techniques with Latent Dirichlet Allocation (LDA) to extract latent topics from raw textual data. These topics are subsequently interpreted through expert judgment, supported by human reviewers and LLM-based agents, in order to identify privacy-sensitive themes and assign privacy risk levels to individual documents. **Results:** The results indicate that the proposed approach facilitates the identification of privacy-relevant topics and supports the classification of documents according to different risk levels. By combining topic modeling with expert interpretation, the approach offers practical support for dataset screening

and privacy-oriented decision-making within Requirements Engineering (RE). **Conclusion:** The case study demonstrates that topic modeling can be effectively employed as a decision-support mechanism for privacy-aware Requirements Engineering, thereby extending RE practices to the governance of AI training data.

5.2 Introduction to the Case Study 2

From a Requirements Engineering (RE) perspective, particularly with regard to the elicitation, analysis, specification, and validation of privacy-related requirements associated with training data, AI systems demand explicit requirements governing how textual data are collected, screened, processed, anonymized, stored, and reused throughout the software lifecycle. Although privacy, security, confidentiality, and compliance are well-established non-functional concerns, in AI-based systems these considerations extend beyond system functionality to encompass the data used for model training. Belani et al. [206] introduced RE4AI ("RE for AI"), a taxonomy that integrates data, model, and system dimensions with the phases of Requirements Engineering, while explicitly incorporating considerations related to dataset privacy and data safety.

The adoption of AI in real-world applications further intensifies concerns surrounding data privacy. Data employed in AI systems are cleaned, integrated, and processed throughout the lifecycle, and each of these stages may introduce threats to individuals' privacy [207]. Consequently, these processes have a direct impact on the ethical handling, management, and protection of data. Early risk analysis is therefore essential, as privacy issues embedded in raw textual data may propagate to subsequent stages of AI development. In this context, Requirements Engineering (RE) activities must address not only system behavior, but also the constraints and governance mechanisms associated with data selection and preparation. This involves translating legal, ethical, and regulatory requirements into operational practices that guide the use of textual data and support decision-making concerning dataset suitability.

The analysis of sensitive-information topics has demonstrated broad applicability across a range of domains [159]. However, the analysis of documents as sources of information by requirements engineers remains a time-consuming and largely manual activity [208]. Given that AI training datasets may comprise thousands or even millions of documents, exhaustive human inspection becomes impractical. This limitation underscores the need for automated or semi-automated approaches capable of identifying privacy-sensitive content in textual artifacts before such data are incorporated into AI pipelines. In this context, topic modeling techniques have proven effective in automatically uncovering latent topics within large volumes of text [209].

The principal contributions of this case study are twofold: (i) the proposal of an approach that integrates LDA-based topic modeling [210] with expert and AI-assisted interpretation to support the identification of privacy risks in textual data; and (ii) the development of a method for classifying documents according to privacy risk levels, thereby supporting Requirements Engineering (RE) activities such as risk analysis and decision-making concerning the suitability of textual data for AI model development [207].

5.3 Background and Related Work to the Case Study 2

Text classification and topic modeling constitute the foundation, the backbone, for the analysis of large textual corpora [211]. In this context, topic modeling offers a promising approach for classifying documents according to privacy risk levels by enabling the identification of latent semantic structures within extensive text collections. In particular, Latent Dirichlet Allocation (LDA), the most widely adopted topic modeling technique [212], represents documents as probabilistic mixtures of topics, with each topic defined as a probability distribution over words [210, 213]. This probabilistic representation facilitates the identification of underlying themes, including those associated with sensitive or privacy-relevant content.

However, topic extraction alone is insufficient for effective privacy risk assessment. Determining whether a given topic constitutes a privacy concern requires contextual interpretation and domain-specific expertise. Accordingly, approaches that combine computational techniques with human judgment are better suited to support privacy-aware decision-making in Requirements Engineering contexts.

Alternative approaches to textual analysis include manual or automated qualitative analysis [214], lexicon-based methods grounded in predefined keyword lists, supervised classification techniques, and embedding-based clustering. Although these methods involve different trade-offs, LDA remains particularly appropriate in contexts where unsupervised, interpretable, and document-level analysis is required. Its probabilistic structure also supports the derivation of indicators, such as topic distributions, that can inform privacy risk classification.

LDA is a generative probabilistic model that assumes documents are composed of multiple latent topics, while both document–topic and topic–word distributions are governed by Dirichlet priors [215, 216]. Under this framework, each document is represented as a probability distribution over topics, and each topic as a probability distribution over the vocabulary. This structure enables the analysis of document similarity within a multidimensional topic space, in which distance measures such as the Jensen–Shannon Distance

may be employed to assess similarity and to support tasks such as recommendation and clustering [217].

A central challenge in LDA concerns the selection of the number of topics (K), which is not inferred automatically and must instead be established through model selection procedures. In practice, multiple models are typically evaluated using different values of K, with the final choice guided by criteria such as interpretability, coherence, and alignment with the analytical objective [218]. This decision has a direct effect on the quality and practical usefulness of the resulting topics. The interpretability of LDA makes it particularly well suited to applications that require human-in-the-loop analysis. By revealing latent thematic structures, topic models facilitate the annotation and organization of large text corpora, thereby supporting tasks such as information retrieval, classification, and corpus exploration [213]. In the context of privacy-aware Requirements Engineering (RE), this capability can be leveraged to identify privacy-sensitive themes and to support early-stage risk assessment in textual data used for AI training.

In this context, the present case study proposes an LDA-based approach [210] to support privacy-aware Requirements Engineering in the analysis of raw textual data used for AI training [207]. The approach combines topic modeling with expert interpretation, supported by human reviewers and LLM-based AI agents, in order to identify privacy-sensitive content and assign risk levels to documents.

5.3.1 Related works to the Case Study 2

Multi-document summarization has attracted considerable scholarly attention, particularly in approaches that incorporate lexical semantics [219]. In a related line of research, Gambarelli et al. [84] examined the automatic identification and classification of sensitive data through the use of labeled datasets designed to distinguish sensitive from non-sensitive content.

Wang et al. [220] proposed PrivScore, a context-sensitive model for the evaluation of private information that supports the detection of privacy breaches. Similarly, Mao et al. [221] examined the presence of sensitive information in tweets, conceptualizing the problem as a binary classification task. Tillmann et al. [222], in turn, combined XGBoost, LDA, and Generalized Additive Models (GAMs) to investigate the relationship between latent topic structures and privacy sensitivity, demonstrating that topic distributions can be associated with variations in privacy risk.

Alemany et al. [223] proposed a sensitivity-classification framework for information shared on social networks, categorizing data according to varying levels of personal sensitivity, ranging from non-personal to highly sensitive information. Likewise, Löbner et al. [33] emphasized that the detection of privacy-sensitive information through machine

learning remains a promising yet still evolving area of research, while also identifying key challenges and gaps that warrant further investigation.

Hiniduma et al. [224] examine data privacy in the context of artificial intelligence, with particular emphasis on the risk of unauthorized disclosure of personal information. The authors argue that privacy breach assessment constitutes a critical factor in determining data readiness and highlight the need to adopt privacy-by-design principles in order to support the ethical and responsible deployment of AI systems.

Martinelli et al. [225] proposed a framework for the semi-automatic construction of annotated corpora by integrating named entity recognition, transfer learning, word embeddings, and topic extraction techniques. Within this approach, human expertise is required primarily in the final stage, where it is employed to validate and refine the automatically generated data.

Despite these advances, existing approaches offer limited support for the integration of topic modeling with privacy-aware decision-making in Requirements Engineering, particularly in the context of AI training data.

5.4 Study Settings in the Case Study 2

This case study is guided by the following research question: **RQ: How can a semi-automated topic analysis approach based on LDA support Requirements Engineering for privacy protection in raw texts used in AI model training?**

To address this question, the study concentrates on identifying privacy-sensitive topics and translating them into document-level risk indicators to support decision-making in Requirements Engineering.

To structure the literature mapping process, this study adopts the PICOC framework proposed by Petticrew and Roberts [93], which comprises the elements of Population, Intervention, Comparison, Outcome, and Context.

The *Population* comprises raw texts and textual artifacts used as input for AI model training, particularly those that may contain privacy-sensitive information. The *Intervention* refers to a semi-automated approach that combines text preprocessing with LDA-based topic modeling in order to identify latent topics associated with privacy risks. The *Comparison* includes manual inspection, expert-driven analysis, and alternative natural language processing and machine learning techniques for the detection of sensitive content. The *Outcome* encompasses both the identification and interpretation of privacy-sensitive topics and the classification of documents according to privacy risk levels. Finally, the *Context* is situated within Requirements Engineering, with particular emphasis on pri-

vacy protection, data suitability assessment, and governance-related decisions concerning textual data used in AI model development.

Based on the research question and the PICOC elements, we constructed the search string presented below. To avoid ambiguity with an unrelated acronym, the term "linear discriminant analysis" was explicitly excluded from the search.

Search String for the Literature Mapping

```
("latent dirichlet allocation" OR "topic model" OR "topic analysis")
AND ("privacy-sensitive" OR "privacy risk" OR "personal data" OR "personally
identifiable information" OR PII OR "sensitive information")
AND ("raw text" OR "textual data" OR corpus OR corpora OR "training data" OR
"training corpus")
AND ("artificial intelligence" OR "AI model" OR "machine learning" OR "model
training" OR "language model" OR "foundation model")
NOT ("linear discriminant analysis")
```

The search was conducted across four major digital libraries: ACM Digital Library [94], IEEE Xplore [95], Scopus [96], and Web of Science [97]. Considering studies published between 2015 and 2026, the search returned 511 records: 487 from ACM Digital Library, 17 from IEEE Xplore, 5 from Scopus, and 2 from Web of Science. After title and abstract screening, 18 studies were selected and retained as primary studies for the literature mapping.

Study selection was conducted in accordance with predefined inclusion criteria (IC), exclusion criteria (EC), and quality assessment (QA) criteria. Studies were included if they satisfied the conditions specified in Table 1, excluded if they met any criterion listed in Table 2, and subsequently evaluated using the quality assessment checklist presented in Table 3.

To manage the literature review process, we employed Parsifal [100], a free and open-source platform specifically developed to support systematic reviews in software engineering. Parsifal was selected because its workflow is closely aligned with the guidelines proposed by Kitchenham and Charters [92], thereby facilitating collaborative screening, duplicate detection, and the traceability of review activities. The initial set of retrieved studies is available in articles.xlsx, and the final set of selected primary studies is available in articles-selected.xlsx. Figure 5.1 presents a summary of the study selection process, adapted to the context of literature mapping in software engineering.

5.5 Method Proposed in the Case Study 2

Based on the insights derived from the literature, we propose a software-based pipeline as a practical response to the research question. The approach consists of a semi-

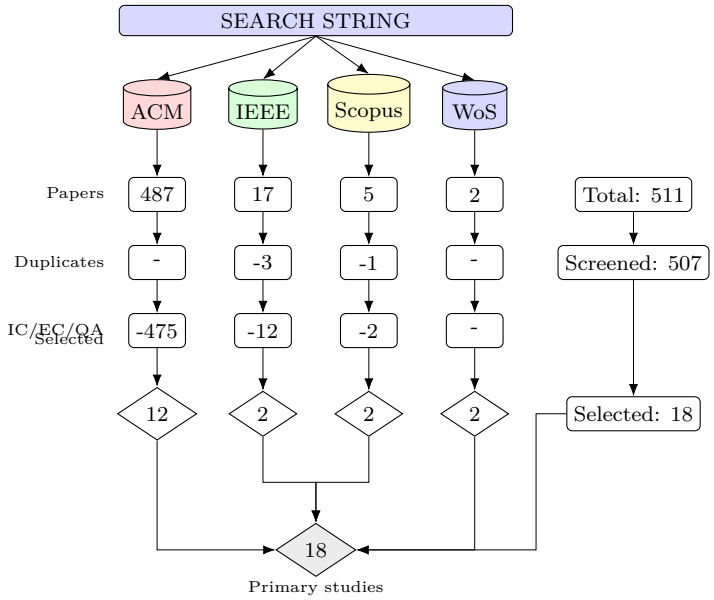


Figure 5.1: Study selection process.

automated topic analysis method, grounded in LDA, to support Requirements Engineering (RE) activities related to privacy protection in raw textual data used for AI model training. The pipeline computes a privacy risk score for each document in a corpus by combining topic modeling with expert interpretation. More specifically, human reviewers and LLM-based AI agents are employed to assess the privacy risk associated with each identified topic. As a result, the approach generates a ranked list of documents with corresponding risk scores, thereby supporting RE decision-making prior to the use of such data for AI model training.

The overall process is structured into three phases: *Preprocessing*, *Training*, and *Scoring*. For the development and initial evaluation of the pipeline, we first employed publicly available datasetss. In a later subsequent, we generated a synthetic dataset with the assistance of ChatGPT (GPT-5.4 Thinking) [226], comprising 1,500 text samples: 1,000 non-sensitive texts and 500 texts containing privacy-sensitive information. The dataset is publicly available on Zenodo at record 19261640.

Preprocessing Phase. Although preprocessing is not the primary focus of this study, it plays a significant role in enhancing the performance of LDA. This phase generally encompasses text collection, cleaning, normalization, and the definition of the unit of analysis (e.g., document, paragraph, or message). Standard text-mining techniques are typically employed, including stopword removal, word-frequency filtering, and stemming or lemmatization [227]. These procedures help reduce noise and consolidate linguistic variation. In addition, tools such as SpaCy [228] may support further linguistic processing,

including part-of-speech tagging.

Training Phase. The training phase constitutes the core of the proposed method. The implementation was developed using the Antigravity IDE [229], with support from LLM-based coding assistants, including Claude Sonnet 4.6 [230] and Gemini 3.1 Pro [231]. During this phase, the LDA model is applied to identify latent topics within the corpus. In our experiments, the number of topics was fixed at $K = 30$, representing a balance between interpretability and granularity. As noted in previous studies, the selection of K is a non-trivial task: lower values may yield excessively broad topics, whereas higher values may compromise interpretability [218]. The implementation relies on the *gensim* library [232, 233, 234, 235]. The resulting topics, together with their most relevant terms, are visualized in Figure 5.2 and subsequently used as input for the labeling process.

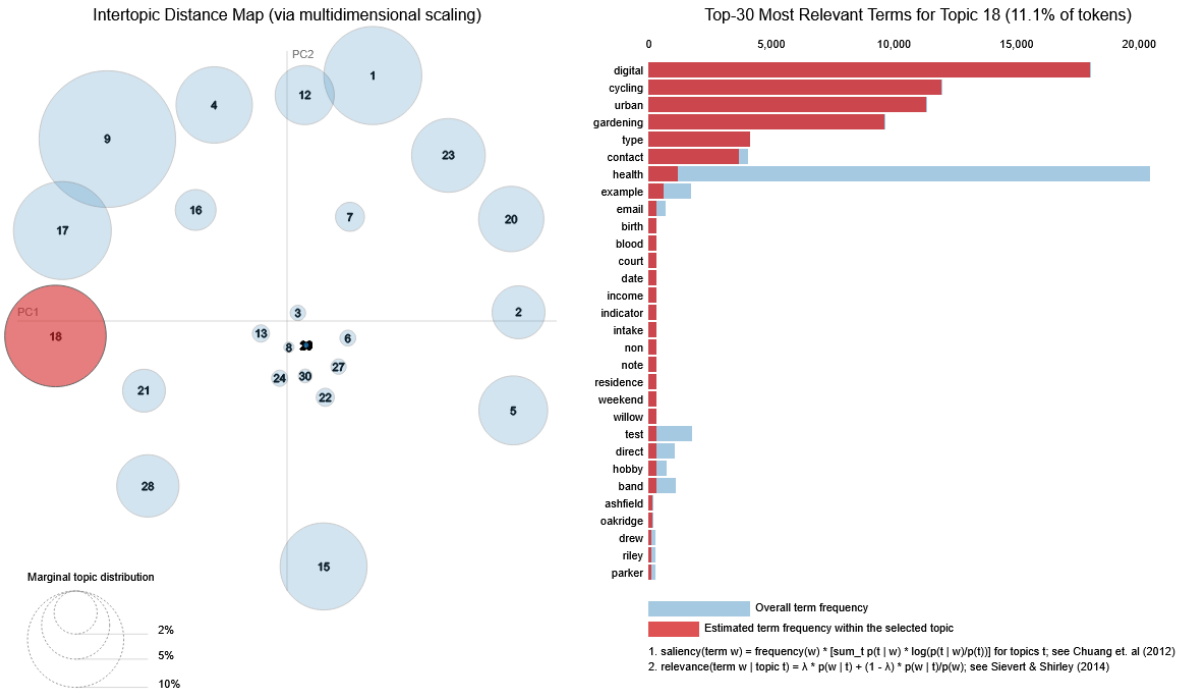


Figure 5.2: Topics distribution and most relevant words per topic

Two alternative labeling workflows are supported. In the **human-centered workflow**, topics are labeled manually by domain experts, who assign to each topic a name, a brief description, and a privacy risk level. In the **AI-assisted workflow**, an LLM automatically generates topic labels, descriptions, and risk levels, which may subsequently be reviewed by human experts if desired. In both workflows, topics are presented through a Topic Labeler interface (Figure 5.3), which enables their interactive inspection and classification.

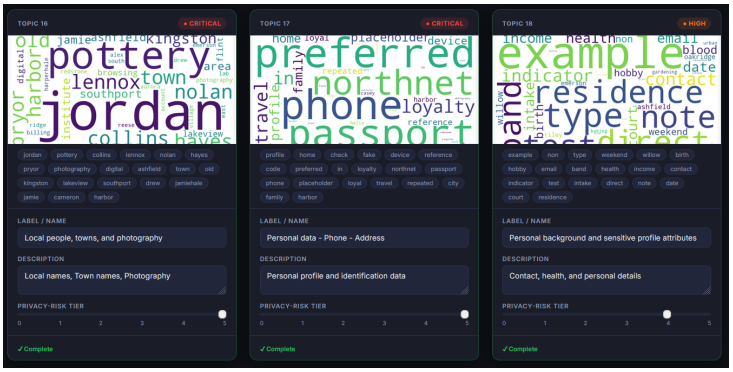


Figure 5.3: Topic Labeler interface

Scoring Phase. Once topics have been labeled according to privacy risk levels, document-level risk scores are computed. Because LDA provides, for each document, a probability distribution over topics (Figure 5.5), the final score is calculated as a weighted average of topic probabilities and their corresponding risk levels. The resulting score ranges from 1 to 5, where 5 denotes the highest level of privacy risk. Figure 5.4 presents an example of the resulting ranked list of documents.

```
Document_ID,Risk_Score,Top_Topic,Text
311,2.9927,23,"Synthetic insurance entry for Jamie Lennox includes policy holder email jamie.lennox311@example.test,
45,2.9925,18,"Example health intake note for Riley Sutton: date of birth 1999-10-19, blood type B-, contact email ril
46,2.9925,16,"Placeholder travel profile for Casey Sutton references passport P-FAKE-700046, preferred phone (555) 01
352,2.991,9,"Fictional commerce account ACCT-FAKE-100352 for Cameron Collins stores username cameroncollins352, billi
41,2.991,28,"Fake profile for Jordan Sutton: born 1995-06-15, lives at 141 Maple Street, Westhaven F51387, can be rea
340,2.991,4,"Test employee file EMP-5340 states that Finley Turner uses badge MBR-20260340, emergency contact number
```

Figure 5.4: Example of document-level privacy risk scores

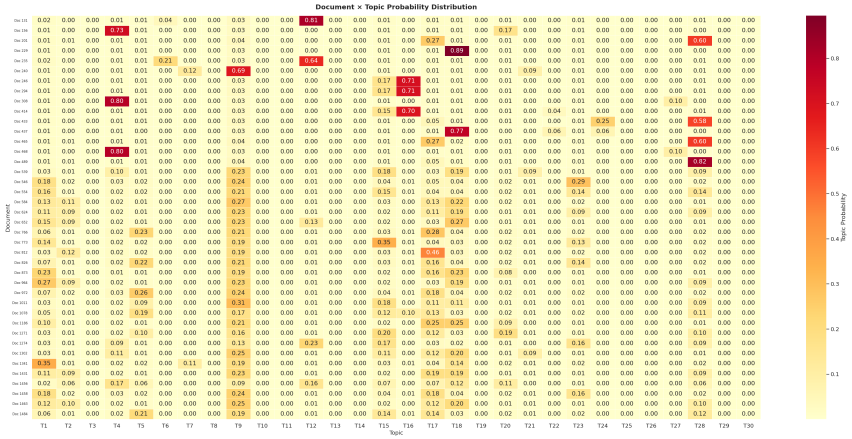


Figure 5.5: Document-topic distribution

The Topic Labeler interface dynamically updates document scores in response to user adjustments, thereby enabling immediate feedback and iterative refinement. The resulting outputs are stored in a structured format (e.g., *topic_labels.json*), which supports traceability and facilitates reuse.

LLM-based Topic Evaluation. In the AI-assisted workflow, a local LLM (Llama 3.1, deployed via Ollama [236, 237]) is employed to evaluate the identified topics. Based on the top keywords associated with each topic, the model generates three outputs: a topic label, a brief description, and a privacy risk tier on a scale from 1 to 5. The adoption of a local deployment strategy was motivated by the need to mitigate privacy risks associated with external APIs and to ensure greater control over data processing.

```
You are an expert in privacy requirements engineering and AI data protection.
Given the following top keywords extracted from a single topic in a privacy-requirements corpus:
{', '.join(keywords)}

Based on these keywords:
1. Provide a concise, descriptive 'name' for this topic (e.g. 'Data Security and Protection', 'User Consent and Rights').
2. Provide a short 'description' of what this topic entails (1-2 sentences).
3. Assign a 'privacy_risk_tier' on a scale of 1 to 5, where 1 is Minimal Risk and 5 is Critical Risk.
4. Assign a 'privacy_risk_label' corresponding to the tier (e.g. Minimal, Low, Moderate, High, Critical).

Output exactly and only a valid JSON object matching this structure:
{{
  "name": "Topic Name",
  "description": "Topic Description",
  "privacy_risk_tier": 2,
  "privacy_risk_label": "Low"
}}
"""
```

Figure 5.6: Prompt used for LLM-based topic evaluation

Response to the Research Question

The proposed method addresses the research question by demonstrating that a semi-automated, LDA-based topic analysis approach can support privacy-aware Requirements Engineering in the analysis of raw textual data used for AI model training. Specifically, the method operates by: (i) extracting latent topics from large textual corpora; (ii) combining human and AI-assisted interpretation to identify privacy-sensitive themes; and (iii) translating topic-level assessments into document-level privacy risk scores. In doing so, the method offers practical support for early-stage dataset screening, privacy risk analysis, and informed decision-making regarding the suitability of textual data for AI model training.

5.6 Discussion

The results suggest that the proposed semi-automated approach constitutes a viable mechanism for identifying privacy-relevant themes within raw textual corpora used for AI training. By combining LDA-based topic modeling with human-assisted interpretation and LLM support, the approach enables the transformation of large volumes of unstructured text into a smaller set of interpretable topics and document-level risk indicators. The experiments further indicate that topic boundaries are not always clearly defined, as

some topics encompass both sensitive and non-sensitive terms, thereby rendering fully automatic privacy classification unreliable. This finding reinforces the importance of human validation when topic models are applied in privacy-sensitive contexts.

From a Requirements Engineering perspective, the proposed workflow is particularly valuable because it enables privacy analysis to be anticipated earlier in the AI lifecycle, prior to the incorporation of data into model training pipelines. In this regard, the approach can support Requirements Engineering activities such as dataset screening, privacy risk analysis, and the validation of data adequacy constraints. Rather than replacing expert judgment, the risk scores generated by the pipeline should be understood as decision-support indicators that assist in prioritizing documents for review and in identifying portions of a dataset that may require additional safeguards, anonymization, or exclusion. In doing so, the approach extends the scope of Requirements Engineering beyond system functionality to include the governance of the training data itself.

This is consistent with prior studies that emphasize the importance of human-in-the-loop approaches for interpreting machine learning outputs in sensitive domains. The present study offers a useful proof of concept and a practical foundation for future research involving real-world, heterogeneous corpora, annotated datasets, alternative topic-modeling techniques, and more systematic empirical validation.

5.6.1 Threats to Validity the Approach from the Case Study 2

As with all empirical and literature-based studies, this work is subject to limitations that may affect the validity and generalizability of its findings. The principal threats to validity and the corresponding mitigation strategies are discussed below, following established guidelines for systematic reviews [92].

Internal Validity: Bias may have been introduced during the study selection and data analysis processes. To mitigate this risk, the literature mapping was conducted according to a predefined protocol, including explicit inclusion and exclusion criteria, a PICOC-guided search strategy, and a quality assessment checklist. The use of Parsifal [100] further supported traceability and consistency throughout the review process. Moreover, the interpretation of topics and the assignment of privacy risk levels inherently involve human judgment, which may introduce subjectivity. This limitation was partially mitigated through the combined use of human and LLM-assisted analysis.

External Validity: The generalizability of the findings is constrained by the fact that the evaluation was conducted using a specific LDA configuration, a synthetic dataset, and a controlled experimental setting. Although these conditions are adequate to demonstrate the feasibility of the proposed approach, further validation is necessary using diverse real-world datasets, particularly annotated corpora, as well as through comparisons with al-

ternative topic-modeling and classification techniques. To support such future extensions, the pipeline was designed in a modular manner, and a hybrid human–AI labeling process was adopted.

Construct Validity: The study operationalizes privacy risk through topic interpretation and document-level scoring derived from LDA outputs. However, privacy risk is an inherently complex and context-dependent construct that may not be fully captured through topic distributions alone. To mitigate this limitation, the proposed approach incorporates expert interpretation and grounds the analysis in privacy-related concepts identified in the literature. Nevertheless, the resulting risk scores should be understood as indicative rather than definitive measures.

Scope-Specific Considerations: This study is specifically concerned with raw textual data used for AI training and with an unsupervised, LDA-based approach for early-stage privacy risk screening. Accordingly, the findings should not be generalized to other data modalities or to tasks that require supervised classification or formal compliance verification. This limitation is addressed by explicitly delimiting the scope of the study and by positioning the proposed method as a decision-support tool. In addition, the adoption of a flexible architecture supports the future integration of alternative models and complementary privacy-preserving techniques.

5.7 Chapter Summary and Future Work from the Case Study 2

Future work should emphasize the strengthening of empirical validation through comparative experiments, hyperparameter optimization, and a broader evaluation of alternative topic-modeling techniques and LLM configurations.

Further investigation into Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic appears particularly promising. NMF is recognized for its computational efficiency and scalability, Top2Vec offers support for multilingual analysis, and BERTopic provides versatility and stability across different domains [238]. Accordingly, these methods constitute relevant alternatives for extending topic-modeling approaches in privacy-aware Requirements Engineering for AI training data.

Another important direction for future research is the evaluation of different LLM models, an objective made feasible by the modular and containerized architecture of the proposed pipeline. This design supports both model replacement and systematic comparison across configurations, thereby enabling more robust experimental validation.

Hyperparameter tuning in LDA also remains a central challenge. Future work may therefore investigate systematic optimization strategies, such as grid search for selecting

the number of topics (K), as suggested in prior studies [218, 239, 240]. The performance of LDA depends not only on the quality of the dataset, but also on parameters such as K, α, and β [218].

In conclusion, this case study demonstrated that the integration of LDA-based topic modeling [210] with human and LLM-assisted interpretation can effectively support privacy-aware Requirements Engineering for AI training data [207]. The proposed workflow translates topic extraction into a practical decision-support mechanism for privacy risk analysis and requirements validation, thereby extending the scope of Requirements Engineering beyond system functionality to encompass data governance. The results indicate that the approach is both feasible and useful as an early-stage mechanism for privacy assessment, although the incorporation of additional privacy-preserving techniques remains an important avenue for future research.

Artifact Availability for the Case Study 2

All source code and supporting materials developed to implement the proposed approach in practice and to substantiate the findings of this study are openly available at https://github.com/Daniel-Lim-Apo/Topic-Modeling-Privacy-for-AI-Training and https://zenodo.org/records/19261640.

Chapter 6

Conclusion

This work propose identify contemporary and state-of-the-art practices in privacy-preserving in texts. A SMS was undertaken with this objective. The results from the SMS were used as a basis for application in a case study exploring methods for rare event detection in an approach integrating dimensionality reduction, vector database, and similarity metrics such as cosine distance to detect semantic outliers and similar texts.

The SMS resulted in 2140 papers and after the protocol conduction a final number of 86 primary studies. As SMS results, we defined a map of state-of-the-art techniques for semantic similarity and rarity analysis in textual data, including methodological classifications and application contexts and a identification of gaps and opportunities for future research about applications of **Large Language Models (LLMs)** and **AI agents** in the processing of sensitive or privacy-critical textual information.

From the SMS we noticed the maturity and relevant presence of text similarity techniques in several contexts, beyond the rare events detection in a differential privacy . Classic methods, like cosine similarity, in the set set of novel approaches as AI Agents.

The case study 1 was planned and executed in methodology with a distributed architecture inside docker containers in two principal flows, the first with dimensionality reduction of the texts by summarization by AI agents with LLM and another flow without summarization and in both with a subsequent process of word embedding and vectorization for cosine distance search approach.

The second case study built upon the findings of the systematic mapping study (SMS), which identified the main techniques employed for semantic similarity and rare-event detection in text data under differential privacy constraints. Among the broad families of approaches observed, topic modeling and probabilistic methods emerged as especially prominent, with Latent Dirichlet Allocation (LDA)[210] appearing in the majority of the reviewed studies. From a Requirements Engineering (RE) perspective, this case study

demonstrated that the integration of LDA-based topic modeling with human and LLM-assisted interpretation can effectively support privacy-aware Requirements Engineering for AI training data. The proposed workflow translates topic extraction into a practical decision-support mechanism for privacy risk analysis and requirements validation, thereby extending the scope of Requirements Engineering beyond system functionality to encompass data governance. Overall, the results indicate that the approach is both feasible and useful as an early-stage mechanism for privacy assessment, although the incorporation of additional privacy-preserving techniques remains an important avenue for future research.

An undesirable but interesting and valuable result in case study to be considered in the future is that the models sometimes refuse to process the data refusing to process the data due to its illegal, immoral, or unethical nature, even though the context of the case is public safety and this content, unusual or inappropriate in other domains, is very necessary in the context of public safety.

The most satisfactory results in case study were obtained when the texts were vectorized in their original form, without prior dimensionality reduction through summarization by AI agents. The process enabled an interactive and easy-to-use user interface for understanding similar, rare, and grouped texts, allowing users to navigate between neighboring texts and perform the same analyses.

In conclusion, this study contributes both a synthesis of current research through a SMS and a practical demonstration through a case study. The study highlights techniques used to preserve privacy in text analysis. The results emphasize the growing role of semantic similarity methods and vector-based representations in detecting rare and sensitive events, particularly in contexts where privacy is required. Moreover, the integration of LLMs and AI agents introduces novel avenues for privacy-aware processing, although with operational challenges such as content moderation and ethical filtering. These insights point to a need, that is growing, for more nuanced frameworks, or the need of to build customized LLM models, that balance technical efficacy with ethical considerations, especially in domains like public security. The approach of this study was not found in many works present in the SMS; therefore, the findings and the consequent stimulus to discussion contribute to filling this gap, comparing them with the few directly related works and serving as inspiration for future research aimed at reducing the distance between advanced AI methodologies and the responsible handling of privacy-critical textual information.

Contributions of this study:

- Systematization of Privacy-Preserving Techniques for Textual Data Privacy research has focused on structured/tabular data, but there were also studies about privacy

risks in textual data. This study provided a taxonomy of privacy-preserving techniques for texts.

- Proposal of a process for semantic rarity detection as a proxy for privacy risk. Rare or unique events in text are hard to detect but pose high reidentification risk. This study proposes using semantic embeddings to represent documents in vector space, than uses distance metrics and outlier detection with semantic isolation in clusters to find texts that are semantically rare and introduces the idea of a privacy rarity score based on vector similarity.

- A process using containerized LLMs and AI Agents. Cloud-based learning management processes may violate data sovereignty and privacy laws. This study brought a workflow running LLMs with fully local execution using open-source models packaged inside containerized Docker environments, ensuring: Full data control; Offline processing; Consistent software dependencies; Isolation from the internet or other systems; Easy deployment across machines or institutions; Reproducibility of results. The contribution here is a technical privacy-preserving approach via local inference pipelines using LLMs within containers, a privacy pipeline that aligns with security best practices and legal requirements for sensitive data.

- The pipeline in case study 1 similarity allowed the institution owner of the data the possiblility to group several groups of police ocurrences that represents real life relevant facts, as criminal modus operandi in crimes ocurred by internet, like scams and others, that are a crescent kind for crime for what is very important an automated process to found relations specially about individuals acting as an organized crime and also when discover group of crimes with similar behavior of the actors.

- The principal contributions from the case study 2 are twofold. First, it proposes an approach that integrates LDA-based topic modeling [210] with expert-driven and AI-assisted interpretation in order to support the identification of privacy risks in textual data. Second, it introduces a method for classifying documents according to levels of privacy risk, thereby supporting Requirements Engineering (RE) activities, particularly risk analysis and decision-making regarding the appropriateness of textual data for AI model development [207].

References

[1] Schäfer, Franziska, Christian Zeiselmair, Jonas Becker, and Heiner Otten: *Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes*. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 190–195, November 2018. https://ieeexplore.ieee.org/document/8691266, visited on 2025-03-24, ISSN: 2159-5119. viii, xvi, 1, 3, 6, 46, 47

[2] Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez: *A Systematic Literature Review on Applying CRISP-DM Process Model*. Procedia Computer Science, 181:526–534, January 2021, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050921002416, visited on 2024-12-05. viii, xvi, 6, 46, 47

[3] Aghasian, Erfan, Saurabh Garg, and James Montgomery: *An automated model to score the privacy of unstructured information—Social media case*. Computers & Security, 92:101778, May 2020, ISSN 0167-4048. https://www.sciencedirect.com/science/article/pii/S0167404820300638, visited on 2025-03-24. 1

[4] Dwork, Cynthia: *Differential Privacy*. In Bugliesi, Michele, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (editors): *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer, ISBN 978-3-540-35908-1. 1, 13, 45

[5] Dwork, Cynthia: *Differential Privacy: A Survey of Results*. In Agrawal, Manindra, Dingzhu Du, Zhenhua Duan, and Angsheng Li (editors): *Theory and Applications of Models of Computation*, volume 4978, pages 1–19. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, ISBN 978-3-540-79227-7 978-3-540-79228-4. http://link.springer.com/10.1007/978-3-540-79228-4_1, visited on 2025-03-10, Series Title: Lecture Notes in Computer Science. 1, 4, 6, 9, 45

[6] Hassan, Muneeb Ul, Mubashir Husain Rehmani, and Jinjun Chen: *Differential Privacy Techniques for Cyber Physical Systems: A Survey*. IEEE Communications Surveys & Tutorials, 22(1):746–789, 2020, ISSN 1553-877X. https://ieeexplore.ieee.org/document/8854247, visited on 2025-04-13. 1, 3

[7] Cui, Jianhao, Hua Shen, and Ying Cao: *Survey on the Applications of Differential Privacy*. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 43–47, December 2024. https://ieeexplore.ieee.org/document/10912963, visited on 2025-04-11. 1, 9, 45

[8] SWEENEY, LATANYA: *k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, May 2012. https://www.worldscientific.com/worldscinet/ijufks, visited on 2025-06-02, Publisher: World Scientific Publishing Company. 1, 44

[9] Lee, Sejong, Yushin Kim, Yongseok Kwon, and Sunghyun Cho: *Secure privacy-preserving record linkage system from re-identification //attack*. PLOS ONE, 20(1):e0314486, January 2025, ISSN 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0314486, visited on 2025-04-11, Publisher: Public Library of Science. 1, 12, 14, 44

[10] Souza, F. C., R. F. Nogueira, and R. A. Lotufo: *BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis*. Applied Soft Computing, 149:110901, December 2023, ISSN 1568-4946. https://www.sciencedirect.com/science/article/pii/S1568494623009195, visited on 2025-03-13. 1, 7, 16, 44

[11] Curzon, James, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib: *Privacy and Artificial Intelligence*. IEEE Transactions on Artificial Intelligence, 2(2):96–108, April 2021, ISSN 2691-4581. https://ieeexplore.ieee.org/document/9450036, visited on 2025-04-08. 1, 7, 44

[12] Acharya, Deepak Bhaskar, Karthigeyan Kuppan, and B. Divya: *Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey*. IEEE Access, 13:18912–18936, 2025, ISSN 2169-3536. https://ieeexplore.ieee.org/document/10849561, visited on 2025-06-04. 1, 7, 15, 18, 44, 48

[13] Duan, Zhihua and Jialin Wang: *Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI*, November 2024. http://arxiv.org/abs/2411.18241, visited on 2025-01-31, arXiv:2411.18241 [cs]. 1, 7, 16, 18, 44, 48

[14] Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian Yun Nie, and Ji Rong Wen: *A Survey of Large Language Models*, March 2025. http://arxiv.org/abs/2303.18223, visited on 2025-04-09, arXiv:2303.18223 [cs]. 1, 7

[15] Anthropic: *Model context protocol*, 2024. https://modelcontextprotocol.io/introduction. Accessed: April 9, 2025. 1, 7

[16] Khoei, Tala Talaei, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei: *A Survey of the Model Context Protocol (MCP): Standardizing Context to Enhance Large Language Models (LLMs)*, April 2025. https://www.preprints.org/manuscript/202504.0245, visited on 2025-04-09. 1, 7

[17] Nelson, Gregory: *Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification*, April 2015. 1, 7, 12, 13, 44

[18] Lison, Pierre, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid: *Anonymisation Models for Text Data: State of the art, Challenges and Future Directions*. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli (editors): *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online, August 2021. Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.323/, visited on 2025-03-07. 1, 5, 7, 12, 13, 31, 44

[19] Giampaolo, Fabio, Stefano Izzo, Edoardo Prezioso, Diletta Chiaro, Salvatore Cuomo, Valerio Bellandi, and Francesco Piccialli: *A Privacy Preserving Service-Oriented Approach for Data Anonymization Through Deep Learning*. In *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 0738–0746, November 2023. https://ieeexplore.ieee.org/document/10361409, visited on 2025-06-02, ISSN: 2837-0740. 1, 7, 12, 13, 31, 44

[20] Giampaolo, Fabio, Stefano Izzo, Stefano Siccardi, Antongiacomo Polimeno, Valerio Bellandi, and Francesco Piccialli: *Real-Time Anonymization of Sensitive Personal Data Using a Service-Based Architecture*. In *2023 IEEE International Conference on Web Services (ICWS)*, pages 701–703, July 2023. https://ieeexplore.ieee.org/document/10248240, visited on 2025-06-02, ISSN: 2836-3868. 1, 7, 12, 13, 44

[21] Raj, Anushree and Rio D'Souza: *Performance Metrics Evaluation Towards The Effectiveness of Data Anonymization*. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–5, April 2023. https://ieeexplore.ieee.org/document/10126310, visited on 2025-06-02. 1, 7, 12, 13, 14, 44

[22] Murin, M., S. Molčan, M. Michalkc, O. Kainz, and D. Cymbalák: *Technical Solutions for the Processing, Management and Anonymisation of Personal Data in Databases According to EU Data Protection Regulations*. In *2024 International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 490–503, October 2024. https://ieeexplore.ieee.org/document/10850866, visited on 2025-06-02. 1, 7, 12, 14, 44

[23] Asimopoulos, Dimitris, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K. Goudos, Ioannis D. Moscholios, Konstantinos E. Psannis, and Panagiotis Sarigiannidis: *Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches*. In *2024 13th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–6, June 2024. https://ieeexplore.ieee.org/document/10615642, visited on 2025-10-18, ISSN: 2993-4443. 1, 7, 14, 31, 39, 44

[24] Shyalika, Chathurangi, Ruwan Wickramarachchi, and Amit Sheth: *A Comprehensive Survey on Rare Event Prediction*, October 2024. http://arxiv.org/abs/2309.11356, visited on 2025-02-27, arXiv:2309.11356 [cs]. 1, 7, 12, 13, 14, 44

[25] Abubakar, Yahaya Idris, Alice Othmani, Patrick Siarry, and Aznul Qalid Md Sabri: *A Systematic Review of Rare Events Detection Across Modalities Using Machine Learning and Deep Learning*. IEEE Access, 12:47091–47109, 2024, ISSN 2169-3536. https://ieeexplore.ieee.org/document/10479512, visited on 2025-04-07. 1, 7, 12, 13, 14, 44

[26] Sprugnoli, Rachele and Sara Tonelli: *Novel event detection and classification for historical texts*. Computational Linguistics, 45(2):229–265, June 2019, ISSN 0891-2017. https://doi.org/10.1162/coli_a_00347, visited on 2025-04-09. 1, 7, 12, 13, 14, 44

[27] Mendes, Ricardo and João P. Vilela: *Privacy-Preserving Data Mining: Methods, Metrics, and Applications*. IEEE Access, 5:10562–10582, 2017, ISSN 2169-3536. https://ieeexplore.ieee.org/document/7950921, visited on 2025-03-05, Conference Name: IEEE Access. 1, 3, 5, 7, 20

[28] El Mestari, Soumia Zohra, Gabriele Lenzini, and Huseyin Demirci: *Preserving data privacy in machine learning systems*. Computers & Security, 137:103605, February 2024, ISSN 0167-4048. https://www.sciencedirect.com/science/article/pii/S0167404823005151, visited on 2025-03-24. 2

[29] Rocha, Lucas Dalle and Edna Dias Canedo: *Optimizing compliance: Comparative study of data laws and privacy frameworks*. Journal of Internet Services and Applications, 16(1):431–452, Jul. 2025. https://journals-sol.sbc.org.br/index.php/jisa/article/view/5247. 2, 7

[30] Union, European: *General data protection regulation - gdpr*. https://gdpr-info.eu/, 2025. Accessed: 2025-03-27. 2, 7, 12, 30

[31] Federative Republic of Brazil: *Lei Geral de Proteção de Dados Pessoais (LGPD) – Lei nº 13.709, de 14 de agosto de 2018*. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm, 2025. Accessed: 2025-03-27. 2, 4, 7, 12, 30

[32] Zhao, Ying and Jinjun Chen: *A survey on differential privacy for unstructured data content*. ACM Comput. Surv., 54(10s):207:1–207:28, September 2022, ISSN 0360-0300. https://doi.org/10.1145/3490237, visited on 2025-03-05. 2, 5, 7, 8, 12, 13, 45

[33] Löbner, Sascha, Welderufael B. Tesfay, Vanessa Bracamonte, and Toru Nakamura: *Systematizing the State of Knowledge in Detecting Privacy Sensitive Information in Unstructured Texts using Machine Learning*. In *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–7, August 2023. https://ieeexplore.ieee.org/document/10320187, visited on 2025-04-11, ISSN: 2643-4202. 2, 7, 8, 13, 62

[34] Niu, Yuqi, Shuo Chen, Nadin Kökciyan, and Weidong Qiu: *Analyzing Social Media Comments to Understand and Detect Privacy Violations*. IEEE Transactions on Computational Social Systems, pages 1–14, 2025, ISSN 2329-924X. https://ieeexplore.ieee.org/document/10820501, visited on 2025-04-11. 2, 7, 8, 13

[35] Vidyalakshmi, B.S., Raymond K. Wong, and Chi Hung Chi: *Privacy Preserving Information Dispersal in Social Networks Based on Disposition to Privacy.* In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 372–377, December 2015. https://ieeexplore.ieee.org/document/7463754/citations, visited on 2025-04-11. 2, 7, 8, 13

[36] Pépin, Ian, Furkan Alaca, and Farhana Zulkernine: *Privacy-Preserving Multi-Party Keyword-Based Classification of Unstructured Text Data.* In *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pages 203–210, April 2024. https://ieeexplore.ieee.org/document/10621516, visited on 2025-04-11, ISSN: 2325-2944. 2, 7, 8, 13

[37] Shahriar, Sakib and Rozita Dara: *PriSM: Privacy-Preserving Social Media Text Processing and Analytics Framework.* In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 59–66, December 2024. https://ieeexplore.ieee.org/document/10917830, visited on 2025-04-11, ISSN: 2375-9259. 2, 7, 8, 13, 44

[38] Saraiva, Juliana and Sérgio Soares: *Privacy and Security Documents for Agile Software Engineering: An Experiment of LGPD Inventory Adoption.* In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–9, October 2023. https://ieeexplore.ieee.org/document/10304806, visited on 2025-03-12. 2, 12

[39] Ahmed, Hadeer, Issa Traore, Sherif Saad, and Mohammad Mamun: *Automated detection of unstructured context-dependent sensitive information using deep learning.* Internet of Things, 16:100444, December 2021, ISSN 2542-6605. https://www.sciencedirect.com/science/article/pii/S254266052100086X, visited on 2025-04-07. 2, 7

[40] Xiong, Jichao, Jiageng Chen, Junyu Lin, Dian Jiao, and Hui Liu: *Enhancing privacy-preserving machine learning with self-learnable activation functions in fully homomorphic encryption.* Journal of Information Security and Applications, 86:103887, November 2024, ISSN 2214-2126. https://www.sciencedirect.com/science/article/pii/S2214212624001893, visited on 2025-03-24. 3, 7

[41] Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong: *Federated Machine Learning: Concept and Applications.* ACM Trans. Intell. Syst. Technol., 10(2):12:1–12:19, January 2019, ISSN 2157-6904. https://doi.org/10.1145/3298981, visited on 2024-12-30. 3, 31

[42] Gutiérrez, Norma, Beatriz Otero, Eva Rodríguez, Gladys Utrera, Sergi Mus, and Ramon Canal: *A Differential Privacy protection-based federated deep learning framework to fog-embedded architectures.* Engineering Applications of Artificial Intelligence, 130:107689, April 2024, ISSN 0952-1976. https://www.sciencedirect.com/science/article/pii/S0952197623018730, visited on 2025-04-07. 3, 31

[43] Carvalho, Tânia, Nuno Moniz, Pedro Faria, and Luís Antunes: *Survey on privacy-preserving techniques for microdata publication*. ACM Comput. Surv., 55(14s):309:1–309:42, July 2023, ISSN 0360-0300. https://dl.acm.org/doi/10.1145/3588765, visited on 2025-04-13. 3

[44] Asif, Hafiz, Sitao Min, Xinyue Wang, and Jaideep Vaidya: *U.S.-U.K. PETs Prize Challenge: Anomaly Detection via Privacy-Enhanced Federated Learning*. IEEE Transactions on Privacy, 1:3–18, 2024, ISSN 2836-208X. https://ieeexplore.ieee.org/document/10507758, visited on 2024-11-18, Conference Name: IEEE Transactions on Privacy. 3, 31

[45] Vithana, Sajani, Martina Cardone, and Flavio P. Calman: *Private Approximate Nearest Neighbor Search for Vector Database Querying*. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 3666–3671, July 2024. https://ieeexplore.ieee.org/document/10619146, visited on 2025-03-12, ISSN: 2157-8117. 3

[46] Lu, Rongxing, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao: *Toward efficient and privacy-preserving computing in big data era*. IEEE Network, 28(4):46–50, July 2014, ISSN 1558-156X. https://ieeexplore.ieee.org/document/6863131, visited on 2025-04-13. 3

[47] Shree, A N Ramya and P Kiran: *Sensitivity Context Aware Privacy Preserving Text Document Summarization*. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1517–1523, November 2020. https://ieeexplore.ieee.org/document/9297415, visited on 2025-03-24. 3, 4

[48] Yao, LiMing, YingJie Ren, Shuang Sheng, Qian Wang, HaiFeng Liu, XinXin Lv, and Bing Xu: *Index Method of Unstructured Data in Power System Based on Improved B+ Tree*. In *2021 International Conference on Wireless Communications and Smart Grid (ICWCSG)*, pages 574–577, August 2021. https://ieeexplore.ieee.org/document/9616561/authors#authors, visited on 2025-04-04. 3

[49] Berthelier, Gaspard, Antoine Boutet, and Antoine Richard: *Toward training NLP models to take into account privacy leakages*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4854–4862, December 2023. https://ieeexplore.ieee.org/document/10386735, visited on 2025-04-08. 4, 6, 13

[50] Kluge Corrêa, Nicholas: *Dynamic Normativity*. Thesis, Universitäts- und Landesbibliothek Bonn, June 2024. https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/11595, visited on 2025-03-13, Accepted: 2024-06-11T12:54:16Z. 4, 12

[51] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI magazine, 17(3):37–37, 1996. 5

[52] Zhou, Shuheng, Katrina Ligett, and Larry Wasserman: *Differential privacy with compression*. In *2009 IEEE International Symposium on Information Theory*, pages 2718–2722, June 2009. https://ieeexplore.ieee.org/document/5205863, visited on 2025-03-12, ISSN: 2157-8117. 5

[53] Volodina, Elena, Simon Dobnik, Therese Lindström Tiedemann, and Xuan Son Vu: *Grandma Karl is 27 years old – research agenda for pseudonymization of research data.* In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 229–233, July 2023. https://ieeexplore.ieee.org/document/10234012, visited on 2025-10-19. 5, 31, 32

[54] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean: *Distributed Representations of Words and Phrases and their Compositionality.* Advances in Neural Information Processing Systems 26 (NIPS 2013), 2013. 6

[55] Bartakke, Dipti, Santosh Kumar, and Aparna Junnarkar: *Text Summarization and Dimensionality Reduction using Learning Approach.* In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–5, November 2020. https://ieeexplore.ieee.org/document/9298250, visited on 2025-06-03. 7, 48

[56] Dwivedi, Yogesh K., Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, Vassilis Galanos, P. Vigneswara Ilavarasan, Marijn Janssen, Paul Jones, Arpan Kumar Kar, Hatice Kizgin, Bianca Kronemann, Banita Lal, Biagio Lucini, Rony Medaglia, Kenneth Le Meunier-FitzHugh, Leslie Caroline Le Meunier-FitzHugh, Santosh Misra, Emmanuel Mogaji, Sujeet Kumar Sharma, Jang Bahadur Singh, Vishnupriya Raghavan, Ramakrishnan Raman, Nripendra P. Rana, Spyridon Samothrakis, Jak Spencer, Kuttimani Tamilmani, Annie Tubadji, Paul Walton, and Michael D. Williams: *Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy.* International Journal of Information Management, 57:101994, April 2021, ISSN 0268-4012. https://www.sciencedirect.com/science/article/pii/S026840121930917X, visited on 2024-12-31. 12

[57] Meszaros, Janos: *The Conflict Between Privacy and Scientific Research in the GDPR.* In *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–6, October 2018. https://ieeexplore.ieee.org/document/8579471/references#references, visited on 2025-03-12. 12

[58] Zadgaonkar, Ashwini and Avinash J. Agrawal: *An Approach for Analyzing Unstructured Text Data Using Topic Modeling Techniques for Efficient Information Extraction.* New Generation Computing, 42(1):109–134, March 2024, ISSN 1882-7055. https://link.springer.com/article/10.1007/s00354-023-00230-5, visited on 2025-04-03, Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer Japan. 12

[59] Lixun, Li, Wang Gaoshan, Dou Zengjie, and Feng Yan: *ATC: An Automatic Text Comparison Tool Based on Diff Algorithm.* In *2020 International Conference on Computer Engineering and Application (ICCEA)*, pages 642–645, March 2020. https://ieeexplore.ieee.org/document/9103831, visited on 2025-06-04. 12, 48

[60] Ries, Lennart, Maximilian Stumpf, Johannes Bach, and Eric Sax: *Semantic Comparison of Driving Sequences by Adaptation of Word Embeddings.* In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages

1–7, September 2020. https://ieeexplore.ieee.org/document/9294364, visited on 2025-06-04. 12, 48

[61] Mhatre, Sonali, Shilpa Satre, Mansi Hajare, Aditi Hire, Aniket Itankar, and Shruti Patil: *Text Comparison Based on Semantic Similarity*. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–5, June 2023. https://ieeexplore.ieee.org/document/10205616, visited on 2025-06-04. 12, 48

[62] Barth, A., A. Datta, J.C. Mitchell, and H. Nissenbaum: *Privacy and contextual integrity: framework and applications*. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15 pp.–198, May 2006. https://ieeexplore.ieee.org/document/1624011, visited on 2025-09-14, ISSN: 2375-1207. 12

[63] Nissenbaum, Helen: *Privacy as Contextual Integrity*. Washington Law Review, 79(1):119, February 2004. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10. 12

[64] Pattakou, Argyri, Vasiliki Diamantopoulou, Christos Kalloniatis, and Stefanos Gritzalis: *A Unified Framework for GDPR Compliance in Cloud Computing*. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, pages 1–9, New York, NY, USA, July 2024. Association for Computing Machinery, ISBN 979-8-4007-1718-5. https://dl.acm.org/doi/10.1145/3664476.3670918, visited on 2024-11-13. 12

[65] Eberendu, Adanma: *Unstructured Data: an overview of the data of Big Data*. International Journal of Computer Trends and Technology, 38:46–50, August 2016. 12

[66] Kulkarni, Poornima and N K Cauvery: *A Big Data Perspective of Individual Privacy Protection Approaches*. In *2021 6th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–5, October 2021. https://ieeexplore.ieee.org/document/9776332, visited on 2025-03-05. 13

[67] Cummings, Rachel, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang: *Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment*. Harvard Data Science Review, 6(1), January 2024, ISSN 2644-2353, 2688-8513. https://hdsr.mitpress.mit.edu/pub/sl9we8gh/release/3, visited on 2025-04-14, Publisher: The MIT Press. 13, 45

[68] Hewage, U. H. W. A., R. Sinha, and M. Asif Naeem: *Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review*. Artificial Intelligence Review, 56(9):10427–10464, September 2023, ISSN 1573-7462. https://link.springer.com/article/10.1007/s10462-023-10425-3, visited on 2025-06-02, Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 9 Publisher: Springer Netherlands. 13

[69] Meyer, Sarina, Pascal Tilli, Pavel Denisov, Florian Lux, Julia Koch, and Ngoc Thang Vu: *Anonymizing Speech with Generative Adversarial Networks to Preserve Speaker Privacy*. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 912–919, January 2023. https://ieeexplore.ieee.org/document/10022601, visited on 2025-09-25. 13, 35, 37, 40, 48

[70] Hu, Yangyu, Haoyu Wang, Tiantong Ji, Xusheng Xiao, Xiapu Luo, Peng Gao, and Yao Guo: *CHAMP: Characterizing Undesired App Behaviors from User Comments Based on Market Policies*. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 933–945, May 2021. ISSN: 1558-1225. 13, 35

[71] Fursov, Ivan, Alexey Zaytsev, Pavel Burnyshev, Ekaterina Dmitrieva, Nikita Klyuchnikov, Andrey Kravchenko, Ekaterina Artemova, Evgenia Komleva, and Evgeny Burnaev: *A Differentiable Language Model Adversarial Attack on Text Classifiers*. IEEE Access, 10:17966–17976, 2022, ISSN 2169-3536. 13, 35

[72] Hu, Jiacheng, Runyuan Bao, Yang Lin, Hanchao Zhang, and Yanlin Xiang: *Accurate Medical Named Entity Recognition Through Specialized NLP Models*. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 578–582, December 2024. https://ieeexplore.ieee.org/document/10912885, visited on 2025-10-02. 13, 14, 32, 33, 34, 35, 37

[73] Vats, Arpita, Zhe Liu, Peng Su, Debjyoti Paul, Yingyi Ma, Yutong Pang, Zeeshan Ahmed, and Ozlem Kalinli: *Recovering from Privacy-Preserving Masking with Large Language Models*. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10771–10775, April 2024. https://ieeexplore.ieee.org/document/10448234, visited on 2025-10-07, ISSN: 2379-190X. 13, 17, 35

[74] Fei, Xinghui, Sheng Chai, Weijie He, Lu Dai, Ruilin Xu, and Lianjin Cai: *A Systematic Study on the Privacy Protection Mechanism of Natural Language Processing in Medical Health Records*. In *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, pages 1819–1824, August 2024. https://ieeexplore.ieee.org/document/10729333, visited on 2025-10-05. 13, 14, 35

[75] Feng, Jun, Laurence T. Yang, Bocheng Ren, Deqing Zou, Mianxiong Dong, and Shunli Zhang: *Tensor Recurrent Neural Network With Differential Privacy*. IEEE Transactions on Computers, 73(3):683–693, March 2024, ISSN 1557-9956. https://ieeexplore.ieee.org/document/10081283, visited on 2025-09-30. 13, 14, 32, 35

[76] Behnia, Rouzbeh, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan: *EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy*. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566, November 2022. ISSN: 2375-9259. 13, 16

[77] Yang, Gaoming, Xinxin Ye, Xianjin Fang, Rongshi Wu, and Li Wang: *Associated Attribute-Aware Differentially Private Data Publishing via Microaggregation*.

IEEE Access, 8:79158–79168, 2020, ISSN 2169-3536. https://ieeexplore.ieee.org/document/9078085, visited on 2025-10-05. 14, 32

[78] Nethravathi, N P, Prasanth G Rao, Vaibhav J Desai, P Deepa Shenoy, K R Venugopal, and M Indiramma: *SWCTE: Semantic weighted context tagging engine for privacy preserving data mining.* In *2016 International Conference on Data Science and Engineering (ICDSE)*, pages 1–5, August 2016. https://ieeexplore.ieee.org/document/7823968, visited on 2025-10-05. 14, 32, 33, 35, 40

[79] Benard Magara, Maake, Sunday O. Ojo, and Tranos Zuva: *A comparative analysis of text similarity measures and algorithms in research paper recommender systems.* In *2018 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–5, March 2018. https://ieeexplore.ieee.org/document/8368766, visited on 2025-09-24. 15, 37, 38, 39, 44, 48

[80] Sitikhu, Pinky, Kritish Pahi, Pujan Thapa, and Subarna Shakya: *A Comparison of Semantic Similarity Methods for Maximum Human Interpretability.* In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–4, November 2019. https://ieeexplore.ieee.org/document/8947433, visited on 2025-09-24. 15, 37, 38, 44, 48

[81] Wang, Jianguo, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, and Charles Xie: *Milvus: A Purpose-Built Vector Data Management System.* In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, pages 2614–2627, New York, NY, USA, June 2021. Association for Computing Machinery, ISBN 978-1-4503-8343-1. https://dl.acm.org/doi/10.1145/3448016.3457550, visited on 2024-11-17. 15

[82] Tang, Huacong: *Study on Multi-Agent Interactive Learning Models Oriented Towards Dynamic Interactive Topologies.* In *2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, pages 271–275, December 2024. https://ieeexplore.ieee.org/document/10869655, visited on 2025-10-07, ISSN: 2831-4549. 15, 35

[83] Chakrabarti, Dipankar, Neelam Patodia, Udayan Bhattacharya, Indranil Mitra, Satyaki Roy, Jayanta Mandi, Nandini Roy, and Prasun Nandy: *Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support.* In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 0683–0688, October 2018. https://ieeexplore.ieee.org/document/8650382, visited on 2025-10-01, ISSN: 2159-3450. 16

[84] Gambarelli, Gaia, Aldo Gangemi, and Rocco Tripodi: *Is Your Model Sensitive? SPEDAC: A New Resource for the Automatic Classification of Sensitive Personal Data.* IEEE Access, 11:10864–10880, 2023, ISSN 2169-3536. 16, 33, 34, 35, 36, 37, 39, 40, 48, 62

[85] Martinelli, Fabio, Fiammetta Marulli, Francesco Mercaldo, Stefano Marrone, and Antonella Santone: *Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence*. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2020. https://ieeexplore.ieee.org/document/9206801, visited on 2025-09-28, ISSN: 2161-4407. 17, 40

[86] Saeed, Muhammad Yahya, Muhammad Awais, Ramzan Talib, and Muhammad Younas: *Unstructured Text Documents Summarization With Multi-Stage Clustering*. IEEE Access, 8:212838–212854, 2020, ISSN 2169-3536. https://ieeexplore.ieee.org/document/9269966, visited on 2025-10-01. 17, 35, 37, 38, 39, 40, 44, 48, 49, 54

[87] Vinaykumar, Kotte, Srinivasan Rajavelu, and R Elijah Blessing: *Similarity Measures and Text Documents Classfication Accuracies Using Benchmark Datasets*. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 688–695, June 2020. https://ieeexplore.ieee.org/document/9142892, visited on 2025-05-03. 17, 46

[88] Lai, Shuzhong and Dian Lei: *Calculation of sentence vector similarity based on fasttext model of weighted fusion*. In *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pages 1–6, April 2022. https://ieeexplore.ieee.org/document/9849804, visited on 2025-05-04. 17

[89] Huang, Guimin and Canjun Shi: *Short Text Similarity Model Using Semantic and Word Order Features*. In *2024 4th International Conference on Computer Science and Blockchain (CCSB)*, pages 423–427, September 2024. https://ieeexplore.ieee.org/document/10735539, visited on 2025-05-04. 17

[90] Zhang, Jianqiang, Fangxu Wang, Futan Ma, and Guoxing Song: *Text Similarity Calculation Method Based on Optimized Cosine Distance*. In *2022 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 37–39, September 2022. https://ieeexplore.ieee.org/document/9997046, visited on 2025-05-04. 18

[91] Chandrasekaran, Muthukumarapandian: *Enhancing Efficiency and Flexibility of Rapid Prototyping for Scalable Multimodal Intelligent Agents*. In *2024 Artificial Intelligence for Business (AIxB)*, pages 66–71, December 2024. https://ieeexplore.ieee.org/document/10771289, visited on 2025-05-04. 18, 48

[92] Barbara, Kitchenham and Stuart Charters: *Guidelines for performing systematic literature reviews in software engineering*. Keele University, UK, 9:1–65, 2007. 19, 26, 41, 64, 69

[93] Petticrew, Mark and Helen Roberts: *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008. 21, 63

[94] Association for Computing Machinery: *ACM Digital Library*. https://dl.acm.org/. Accessed: 2025-09-15. 24, 64

[95] Institute of Electrical and Electronics Engineers: *IEEE Xplore Digital Library*. https://ieeexplore.ieee.org/. Accessed: 2025-09-15. 24, 64

[96] Elsevier: *Scopus*. https://www.scopus.com/. Accessed: 2025-09-15. 24, 64

[97] Clarivate: *Web of Science*. https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/. Accessed: 2025-09-15. 24, 64

[98] Brereton, Pearl, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil: *Lessons from applying the systematic literature review process within the software engineering domain*. Journal of Systems and Software, 80(4):571–583, April 2007, ISSN 01641212. https://linkinghub.elsevier.com/retrieve/pii/S016412120600197X, visited on 2025-09-15. 24

[99] Merrouni, Zakariae Alami, Bouchra Frikh, and Brahim Ouhbi: *Automatic keyphrase extraction: An overview of the state of the art*. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 306–313, October 2016. https://ieeexplore.ieee.org/document/7805062, visited on 2025-09-26, ISSN: 2327-1884. 24

[100] Parsifal Developers: *Parsifal: A platform for formal modeling and verification of privacy-preserving systems*. https://parsif.al, 2025. Accessed: 2025-09-21. 26, 64, 69

[101] Hassan, Fadi, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer: *Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings*. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365, August 2019. https://ieeexplore.ieee.org/document/8887419, visited on 2025-10-18, ISSN: 2324-9013. 30, 31, 32, 37, 48

[102] McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas: *Communication-Efficient Learning of Deep Networks from Decentralized Data*. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017. https://proceedings.mlr.press/v54/mcmahan17a.html, visited on 2025-01-06, ISSN: 2640-3498. 31

[103] Khan, Younas, David Sánchez, and Josep Domingo-Ferrer: *Federated learning-based natural language processing: a systematic literature review*. Artificial Intelligence Review, 57(12):320, October 2024, ISSN 1573-7462. https://doi.org/10.1007/s10462-024-10970-5, visited on 2025-04-03. 31

[104] Fu, Jie, Yuan Hong, Xinpeng Ling, Leixia Wang, Xun Ran, Zhiyu Sun, Wendy Hui Wang, Zhili Chen, and Yang Cao: *Differentially Private Federated Learning: A Systematic Review*, May 2024. http://arxiv.org/abs/2405.08299, visited on 2025-04-13, arXiv:2405.08299 [cs]. 31

[105] Liu, Yi, James J. Q. Yu, Jiawen Kang, Dusit Niyato, and Shuyu Zhang: *Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach*. IEEE Internet of Things Journal, 7(8):7751–7763, August 2020, ISSN 2327-4662. https://ieeexplore.ieee.org/document/9082655, visited on 2025-09-29. 31, 32, 35

[106] Lim, Wei Yang Bryan, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao: *Federated Learning in Mobile Edge Networks: A Comprehensive Survey*. IEEE Communications Surveys & Tutorials, 22(3):2031–2063, 2020, ISSN 1553-877X. https://ieeexplore.ieee.org/document/9060868, visited on 2024-12-31, Conference Name: IEEE Communications Surveys & Tutorials. 31

[107] Tan, Alysa Ziying, Han Yu, Lizhen Cui, and Qiang Yang: *Towards Personalized Federated Learning*. IEEE Transactions on Neural Networks and Learning Systems, 34(12):9587–9603, December 2023, ISSN 2162-2388. https://ieeexplore.ieee.org/document/9743558, visited on 2025-01-02, Conference Name: IEEE Transactions on Neural Networks and Learning Systems. 31

[108] Wei, Kang, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor: *Federated Learning With Differential Privacy: Algorithms and Performance Analysis*. IEEE Transactions on Information Forensics and Security, 15:3454–3469, 2020, ISSN 1556-6021. https://ieeexplore.ieee.org/document/9069945, visited on 2024-12-31, Conference Name: IEEE Transactions on Information Forensics and Security. 31

[109] Guo, Tao, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu: *PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models – Federated Learning in Age of Foundation Model*. IEEE Transactions on Mobile Computing, 23(5):5179–5194, May 2024, ISSN 1558-0660. https://ieeexplore.ieee.org/document/10210127, visited on 2025-09-29. 31, 32, 40

[110] Yan, Biwei, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng: *On Protecting the Data Privacy of Large Language Models (LLMs): A Survey*. In *2024 International Conference on Meta Computing (ICMC)*, pages 1–12, June 2024. https://ieeexplore.ieee.org/document/11062758, visited on 2025-10-06. 32, 34, 36

[111] Singhofer, F., A. Garifullina, M. Kern, and A. Scherp: *A novel approach on the joint de-identification of textual and relational data with a modified mondrian algorithm*. In *Proceedings of the 21st ACM Symposium on Document Engineering*, DocEng '21, pages 1–10, New York, NY, USA, August 2021. Association for Computing Machinery, ISBN 978-1-4503-8596-1. https://doi.org/10.1145/3469096.3469871, visited on 2025-10-20. 32, 35, 40

[112] Weggenmann, Benjamin, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum: *DP-VAE: Human-Readable Text Anonymization for Online Reviews with Differentially Private Variational Autoencoders*. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 721–731, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 978-1-4503-9096-5. https://doi-org.

ez54.periodicos.capes.gov.br/10.1145/3485447.3512232, event-place: Virtual Event, Lyon, France. 32

[113] Staalduine, Nina van and Anneke Zuiderwijk: *Exploring the Viability of ChatGPT for Personal Data Anonymization in Government: A Comprehensive Analysis of Possibilities, Risks, and Ethical Implications*. Digit. Gov.: Res. Pract., 6(2), June 2025. https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3678264, Place: New York, NY, USA Publisher: Association for Computing Machinery. 32, 36

[114] Staufer, Dimitri, Frank Pallas, and Bettina Berendt: *Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification*. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 733–745, New York, NY, USA, June 2024. Association for Computing Machinery, ISBN 979-8-4007-0450-5. https://dl.acm.org/doi/10.1145/3630106.3658936, visited on 2025-10-21. 32, 34

[115] Chevrier, Raphaël, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis: *Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review*. Journal of Medical Internet Research, 21(5):e13484, May 2019. https://www.jmir.org/2019/5/e13484, visited on 2025-10-23, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. 32

[116] Henriksson, Aron, Maria Kvist, and Hercules Dalianis: *Detecting Protected Health Information in Heterogeneous Clinical Notes*. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 393–397. IOS Press, 2017. https://ebooks.iospress.nl/doi/10.3233/978-1-61499-830-3-393, visited on 2025-10-26. 32

[117] Gupta, Brij B., Akshat Gaurav, Varsha Arya, Wadee Alhalabi, Dheyaaldin Alsalman, and P. Vijayakumar: *Enhancing user prompt confidentiality in Large Language Models through advanced differential encryption*. Computers and Electrical Engineering, 116:109215, May 2024, ISSN 0045-7906. https://www.sciencedirect.com/science/article/pii/S0045790624001435, visited on 2025-10-26. 32, 34, 35, 36

[118] Abdalla, Mohamed, Moustafa Abdalla, Graeme Hirst, and Frank Rudzicz: *Exploring the Privacy-Preserving Properties of Word Embeddings: Algorithmic Validation Study*. JOURNAL OF MEDICAL INTERNET RESEARCH, 22(7), July 2020, ISSN 1438-8871. Place: 130 QUEENS QUAY E, STE 1102, TORONTO, ON M5A 0P6, CANADA Publisher: JMIR PUBLICATIONS, INC Type: Article. 32, 35, 37, 48

[119] Meystre, Stéphane M., Óscar Ferrández, F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore: *Text de-identification for privacy protection: A study of its impact on clinical text information content*. Journal of Biomedical Informatics, 50:142–150, August 2014, ISSN 1532-0464. https://www.sciencedirect.com/science/article/pii/S1532046414000136, visited on 2025-10-26. 32

[120] Cardinal, Rudolf N.: *Clinical records anonymisation and text extraction (CRATE): An open-source software system.* BMC Medical Informatics and Decision Making, 17(1), 2017, ISSN 1472-6947. Publisher: BioMed Central Ltd. 32

[121] Demner-Fushman, Dina, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald: *Preparing a collection of radiology examinations for distribution and retrieval.* Journal of the American Medical Informatics Association, 23(2):304–310, March 2016, ISSN 1067-5027. https://doi.org/10.1093/jamia/ocv080, visited on 2025-10-23. 32

[122] Obeid, Jihad S., Paul M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Christine M. Carr, Kevin Gagnon, Tami Crawford, and Stephane M. Meystre: *Impact of De-Identification on Clinical Text Classification Using Traditional and Deep Learning Classifiers.* In OhnoMachado, L and B Seroussi (editors): *MEDINFO 2019: HEALTH AND WELLBEING E-NETWORKS FOR ALL*, volume 264 of *Studies in Health Technology and Informatics*, pages 283–287, NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM, NETHERLANDS, 2019. IOS PRESS, ISBN 978-1-64368-003-3 978-1-64368-002-6. Backup Publisher: French Assoc Med Informat ISSN: 0926-9630 Type: Proceedings Paper. 32, 35

[123] Mehta, Brijesh, Udai Pratap Rao, Ruchika Gupta, and Mauro Conti: *Towards privacy preserving unstructured big data publishing.* Journal of Intelligent & Fuzzy Systems, 36(4):3471–3482, April 2019, ISSN 1064-1246. https://doi.org/10.3233/JIFS-181231, visited on 2025-10-26, Publisher: SAGE Publications. 32, 34, 35, 39, 40

[124] Languré, Alejandro de León and Mahdi Zareei: *Privacy-Preserving Emotion Detection: Evaluating the Trade-Off Between K-Anonymity and Model Performance.* IEEE Access, 13:105901–105910, 2025, ISSN 2169-3536. https://ieeexplore.ieee.org/document/11031447, visited on 2025-10-18. 32, 34, 35

[125] Hadj Ahmed, Bouarara, Abdelmalek Amine, and Hamou Reda Mohamed: *New Private Information Retrieval Protocol Using Social Bees Lifstyle over Cloud Computing.* In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 161–165, February 2015. https://ieeexplore.ieee.org/document/7078687, visited on 2025-10-05. 32

[126] Yang, Zhen, Jiliang Tang, and Huan Liu: *Cloud Information Retrieval: Model Description and Scheme Design.* IEEE Access, 6:15420–15430, 2018, ISSN 2169-3536. https://ieeexplore.ieee.org/document/8272322, visited on 2025-10-06. 32, 33

[127] Aldeen, Yousra Abdul Alsahib S., Mazleena Salleh, and Yazan Aljeroudi: *An innovative privacy preserving technique for incremental datasets on cloud computing.* Journal of Biomedical Informatics, 62:107–116, August 2016, ISSN 1532-0464. https://www.sciencedirect.com/science/article/pii/S1532046416300545, visited on 2025-10-22. 32

[128] Bo, Haohan, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal: *ER-AE: Differentially Private Text Generation for Authorship Anonymization*. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (editors): *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online, June 2021. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.314/, visited on 2025-10-23. 32

[129] Lit, Zhengyang, Shijing Sit, Jianzong Wang, and Jing Xiao: *Federated Split BERT for Heterogeneous Text Classification*. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2022. ISSN: 2161-4407. 32, 34, 35

[130] Han, Chengzhuo, Tingting Yang, Zhengqi Cui, and Xin Sun: *A Privacy-Preserving and Trustworthy Inference Framework for LLM-IoT Integration via Hierarchical Federated Collaborative Computing*. IEEE Internet of Things Journal, pages 1–1, 2025, ISSN 2327-4662. https://ieeexplore.ieee.org/document/11053989, visited on 2025-10-17. 32, 35, 40

[131] Stanik, Christoph, Tim Pietz, and Walid Maalej: *Unsupervised Topic Discovery in User Comments*. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 150–161, September 2021. https://ieeexplore.ieee.org/document/9604745, visited on 2025-10-01, ISSN: 2332-6441. 34, 37, 38, 39, 40, 48, 49, 54

[132] Zhang, Jian, Hai Zhao, and Bao Liang Lu: *A comparative study on two large-scale hierarchical text classification tasks' solutions*. In *2010 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3275–3280, July 2010. https://ieeexplore.ieee.org/document/5580696, visited on 2025-11-13, ISSN: 2160-1348. 34, 38

[133] Nguyen, Nhung T. H., Makoto Miwa, Yoshimasa Tsuruoka, and Satoshi Tojo: *Identifying synonymy between relational phrases using word embeddings*. Journal of Biomedical Informatics, 56:94–102, August 2015, ISSN 1532-0464. https://www.sciencedirect.com/science/article/pii/S1532046415000908, visited on 2025-11-14. 35, 37, 38

[134] Hu, Xuesong, Huajun Zhang, and Youjun Sun: *Chinese Medical Short Text Matching Model Based on Fine-Tuning BERT-Attention-BiLSTM*. In *2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS)*, pages 91–96, June 2023. 35

[135] Al-Ghuribi, Sumaia Mohammed, Shahrul Azman Mohd Noah, and Sabrina Tiun: *Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews*. IEEE Access, 8:218592–218613, 2020, ISSN 2169-3536. https://ieeexplore.ieee.org/document/9279217, visited on 2025-10-01. 35, 38, 39, 40

[136] Ding, Yucheng, Yangwenjian Tan, Xiangyu Liu, Chaoyue Niu, Fandong Meng, Jie Zhou, Ning Liu, Fan Wu, and Guihai Chen: *Personalized Language Model*

Learning on Text Data Without User Identifiers. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pages 224–235, New York, NY, USA, July 2025. Association for Computing Machinery, ISBN 979-8-4007-1245-6. https://dl.acm.org/doi/10.1145/3690624.3709211, visited on 2025-10-21. 35, 40

[137] Joshi, Himanshu Sanjay and Hamed Taherdoost: *Ethics in Natural Language Processing: Addressing Bias, Privacy, and Misinformation.* In *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, pages 359–364, February 2025. https://ieeexplore.ieee.org/document/10941029, visited on 2025-10-07. 35

[138] Ismail, Sabir and M. Shahidur Rahman: *Bangla word clustering based on N-gram language model.* In *2014 International Conference on Electrical Engineering and Information & Communication Technology*, pages 1–5, April 2014. https://ieeexplore.ieee.org/document/6919083, visited on 2025-09-26. 35

[139] Agarwal, Mayank, Ritika Kalia, Vedant Bahel, and Achamma Thomas: *AutoEval: A NLP Approach for Automatic Test Evaluation System.* In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 1–6, September 2021. https://ieeexplore.ieee.org/document/9573769, visited on 2025-09-25. 35

[140] Osman, Ahmed Hamza and Omar Mohammed Barukub: *Graph-Based Text Representation and Matching: A Review of the State of the Art and Future Challenges.* IEEE Access, 8:87562–87583, 2020, ISSN 2169-3536. https://ieeexplore.ieee.org/document/9088989, visited on 2025-09-28. 35

[141] Gupta, Hritvik and Mayank Patel: *Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert.* In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 511–517, March 2021. https://ieeexplore.ieee.org/document/9395976, visited on 2025-09-29. 35

[142] Sultana Ritu, Zakia, Nafisa Nowshin, Md Mahadi Hasan Nahid, and Sabir Ismail: *Performance Analysis of Different Word Embedding Models on Bangla Language.* In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5, September 2018. https://ieeexplore.ieee.org/document/8554681, visited on 2025-09-29. 35, 36

[143] Şenel, Lütfi Kerem, İhsan Utlu, Veysel Yücesoy, Aykut Koç, and Tolga Çukur: *Semantic Structure and Interpretability of Word Embeddings.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10):1769–1779, October 2018, ISSN 2329-9304. https://ieeexplore.ieee.org/document/8364606, visited on 2025-09-29. 35, 37, 38, 39, 40, 48

[144] Zhu, Luyao, Wei Li, Yong Shi, and Kun Guo: *SentiVec: Learning Sentiment-Context Vector via Kernel Optimization Function for Sentiment Analysis.* IEEE Transactions on Neural Networks and Learning Systems, 32(6):2561–2572, June

2021, ISSN 2162-2388. https://ieeexplore.ieee.org/document/9142399, visited on 2025-09-30. 35, 37, 38, 40, 48

[145] Zaware, Sarika, Deep Patadiya, Abhishek Gaikwad, Sanket Gulhane, and Akash Thakare: *Text Summarization using TF-IDF and Textrank algorithm*. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1399–1407, June 2021. https://ieeexplore.ieee.org/document/9453071, visited on 2025-09-30. 35, 38, 40, 44

[146] P., Sunilkumar and Athira P. Shaji: *A Survey on Semantic Similarity*. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–8, December 2019. 35, 37, 38, 48

[147] Mcconkey, Ryan and Oluwafemi Olukoya: *Runtime and Design Time Completeness Checking of Dangerous Android App Permissions Against GDPR*. IEEE Access, 12:1–22, 2024, ISSN 2169-3536. 35

[148] Baal, Simon T. van, Piotr Bogdanski, Araanya Daryanani, Lukasz Walasek, and Philip Newall: *The lived experience of gambling-related harm in natural language*. Psychology of Addictive Behaviors, 39(4):397–409, 2025, ISSN 1939-1501. Place: US Publisher: American Psychological Association. 35, 37, 39, 48

[149] Shahmirzadi, Omid, Adam Lugowski, and Kenneth Younge: *Text Similarity in Vector Space Models: A Comparative Study*. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 659–666, December 2019. https://ieeexplore.ieee.org/document/8999168, visited on 2025-09-30. 37, 38, 44, 48

[150] Qin, Kyle Kai, Shengzhu Wang, Xiaoyan Wei, Yueyan Huang, and Liwei Su: *An Algorithm for Integrating Multimodal Similar Text Resources Based on Large Language Models*. In *2025 7th International Conference on Information Science, Electrical and Automation Engineering (ISEAE)*, pages 612–615, April 2025. 37, 39

[151] Wang, Yu, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu, and Yichen Wu: *A Method of Relation Extraction Using Pre-training Models*. In *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, pages 176–179, December 2020. https://ieeexplore.ieee.org/document/9325805, visited on 2025-11-14, ISSN: 2473-3547. 37, 39

[152] Devine, Peter, Yun Sing Koh, and Kelly Blincoe: *Evaluating Unsupervised Text Embeddings on Software User Feedback*. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 87–95, September 2021. https://ieeexplore.ieee.org/document/9582373, visited on 2025-09-28. 37, 48

[153] Wang, Yuan, Jie Liu, Yalou Huang, and Xia Feng: *Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs*. IEEE Transactions on Knowledge and Data Engineering, 28(7):1919–1933, July 2016, ISSN 1558-2191. https://ieeexplore.ieee.org/document/7412726, visited on 2025-10-01. 37, 38, 39, 40, 48, 49, 54

[154] Wang, Xinzhi, Hui Zhang, and Yi Liu: *Sentence Vector Model Based on Implicit Word Vector Expression*. IEEE Access, 6:17455–17463, 2018, ISSN 2169-3536. https://ieeexplore.ieee.org/document/8325266, visited on 2025-09-30. 37, 40, 48

[155] Lam, An Ngoc, Anh Tuan Nguyen, Hoan Anh Nguyen, and Tien N. Nguyen: *Bug Localization with Combination of Deep Learning and Information Retrieval*. In *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*, pages 218–229, May 2017. 37, 38, 48

[156] Eminagaoglu, Mete and Yilmaz Goksen: *A New Similarity Measure for Document Classification and Text Mining*. In Polychronidou, P, A Horobet, and A Karasavvoglou (editors): *ECONOMIES OF THE BALKAN AND EASTERN EUROPEAN COUNTRIES*, KnE Social Sciences, pages 353–366, OFFICE 4402, X2 TOWER, JLT, PO BOX 488239, DUBAI, 00000, U ARAB EMIRATES, 2020. KNOWLEDGE E. Backup Publisher: Int Hellen Univ, Dept Accounting & Finance; Bucharest Univ Econ Studies, Fac Int Business & Econ; Bucharest Univ Econ Studies, Ctr Res Int Business & Econ; Romanian Acad, Inst Econ Forecasting; Romanian Acad, Ctr Financial & Monetary Res Victor Slavescu ISSN: 2518-668X Type: Proceedings Paper. 37, 38, 39, 44, 48

[157] Roul, Rajendra Kumar, Jajati Keshari Sahoo, and Kushagr Arora: *Modified TF-IDF Term Weighting Strategies for Text Categorization*. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–6, December 2017. https://ieeexplore.ieee.org/document/8487593, visited on 2025-09-29, ISSN: 2325-9418. 37, 38

[158] Abdalla, Mohamed, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirst: *Using word embeddings to improve the privacy of clinical notes*. Journal of the American Medical Informatics Association, 27(6):901–907, 2020, ISSN 1067-5027. Publisher: Oxford University Press. 37, 48

[159] Xu, Guixian, Xu Wu, Haishen Yao, Fan Li, and Ziheng Yu: *Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model*. IEEE Access, 7:21527–21538, 2019, ISSN 2169-3536. https://ieeexplore.ieee.org/document/8633906, visited on 2025-09-29. 37, 38, 60

[160] Alhawarat, M. and M. Hegazi: *Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents*. IEEE Access, 6:42740–42749, 2018, ISSN 2169-3536. https://ieeexplore.ieee.org/document/8402221, visited on 2025-09-29. 38, 39

[161] Poria, Soujanya, Iti Chaturvedi, Erik Cambria, and Federica Bisio: *Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis*. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4465–4473, July 2016. https://ieeexplore.ieee.org/document/7727784, visited on 2025-09-30, ISSN: 2161-4407. 38, 40

[162] Aiello, Luca Maria, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes:

Sensing Trending Topics in Twitter. IEEE Transactions on Multimedia, 15(6):1268–1282, October 2013, ISSN 1941-0077. https://ieeexplore.ieee.org/document/6525357, visited on 2025-09-29. 38, 39

[163] Li, Fang, Huiyu Shen, and Tingting He: *Tag-topic model for semantic knowledge acquisition from blogs*. In *2011 7th International Conference on Natural Language Processing and Knowledge Engineering*, pages 221–226, November 2011. 38

[164] Gualberto, Eder S., Rafael T. De Sousa, Thiago P. De B. Vieira, João Paulo C. L. Da Costa, and Cláudio G. Duque: *From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection*. IEEE Access, 8:76368–76385, 2020, ISSN 2169-3536. 38, 39

[165] Froud, H., R. Benslimane, A. Lachkar, and S. Alaoui Ouatik: *Stemming and similarity measures for Arabic Documents Clustering*. In *2010 5th International Symposium On I/V Communications and Mobile Network*, pages 1–4, September 2010. https://ieeexplore.ieee.org/document/5656417, visited on 2025-09-30. 38, 44

[166] Yin, Jie, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power: *Using Social Media to Enhance Emergency Situation Awareness*. IEEE Intelligent Systems, 27(6):52–59, November 2012, ISSN 1941-1294. 38

[167] Bleik, Said, Meenakshi Mishra, Jun Huan, and Min Song: *Text Categorization of Biomedical Data Sets Using Graph Kernels and a Controlled Vocabulary*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(5):1211–1217, September 2013, ISSN 1557-9964. https://ieeexplore.ieee.org/document/6475935, visited on 2025-09-30. 38, 39, 40

[168] Fan, Zexuan and Xiaolong Xu: *APDPk-Means: A New Differential Privacy Clustering Algorithm Based on Arithmetic Progression Privacy Budget Allocation*. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1737–1742, August 2019. https://ieeexplore.ieee.org/document/8855385, visited on 2025-10-02. 39, 40, 49, 54

[169] Guan, Renchu, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang: *Text Clustering with Seeds Affinity Propagation*. IEEE Transactions on Knowledge and Data Engineering, 23(4):627–637, April 2011, ISSN 1558-2191. https://ieeexplore.ieee.org/document/5560648, visited on 2025-09-30. 39, 40, 49, 54

[170] Chen, Yuanyuan, Yisheng Lv, Xiao Wang, Lingxi Li, and Fei Yue Wang: *Detecting Traffic Information From Social Media Texts With Deep Learning Approaches*. IEEE Transactions on Intelligent Transportation Systems, 20(8):3049–3058, August 2019, ISSN 1558-0016. 39

[171] Tarek, Ahmed, Marwa Mahmoud, Basma Afifi, Maggie Mashaly, and Mervat Abu-Elkheir: *Query-Based Topic Modeling and Trend Analysis in Scientific Literature*. In *2024 International Conference on Microelectronics (ICM)*, pages 1–6, December 2024. ISSN: 2159-1679. 39

[172] Shih, Chin Hong, Bi Cheng Yan, Shih Hung Liu, and Berlin Chen: *Investigating Siamese LSTM networks for text categorization*. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 641–646, December 2017. https://ieeexplore.ieee.org/document/8282104, visited on 2025-09-28. 39

[173] Yu, Zheng, Haixun Wang, Xuemin Lin, and Min Wang: *Understanding Short Texts through Semantic Enrichment and Hashing*. IEEE Transactions on Knowledge and Data Engineering, 28(2):566–579, February 2016, ISSN 1558-2191. https://ieeexplore.ieee.org/document/7286811, visited on 2025-10-01. 39

[174] Hua, Wen, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou: *Short text understanding through lexical-semantic analysis*. In *2015 IEEE 31st International Conference on Data Engineering*, pages 495–506, April 2015. https://ieeexplore.ieee.org/document/7113309, visited on 2025-09-30, ISSN: 2375-026X. 39

[175] Hua, Wen, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou: *Understand Short Texts by Harvesting and Analyzing Semantic Knowledge*. IEEE Transactions on Knowledge and Data Engineering, 29(3):499–512, March 2017, ISSN 1558-2191. https://ieeexplore.ieee.org/document/7476863, visited on 2025-10-01. 39

[176] Liang, Wenxin, Ran Feng, Xinyue Liu, Yuangang Li, and Xianchao Zhang: *GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts*. IEEE Access, 6:43612–43621, 2018, ISSN 2169-3536. 39

[177] Cheng, Jian: *Long Text Topic Mining and Clustering Analysis Base on Doc2Vec-LDA and K-Means*. In *2024 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 9–13, September 2024. https://ieeexplore.ieee.org/document/10935191, visited on 2025-05-28, ISSN: 2160-1348. 44

[178] Narayanan, Arvind and Vitaly Shmatikov: *Robust De-anonymization of Large Sparse Datasets*. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008. https://ieeexplore.ieee.org/document/4531148, visited on 2025-05-16, ISSN: 2375-1207. 44, 45

[179] Sweeney, Latanya: *Weaving Technology and Policy Together to Maintain Confidentiality*. Journal of Law, Medicine & Ethics, 25(2-3):98–110, 1997, ISSN 1073-1105, 1748-720X. https://www.cambridge.org/core/product/identifier/S1073110500005817/type/journal_article, visited on 2025-05-16. 45

[180] AI, Napkin: *Napkin AI: Visual Intelligence for Ideas*. https://www.napkin.ai/, 2025. Accessed: 2025-06-06. 46, 47, 50, 51

[181] Azhar, Muhammad Helmi, I Wayan Widi Pradnyana, and Nindy Irzavika: *Optimization of Microservice-Based Academic Services with the Use of Message Brokers (Case Study: Business Process of Submitting Krs)*. In *2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pages 827–832, November 2024. https://ieeexplore.ieee.org/document/10956829, visited on 2025-06-02, ISSN: 2837-5203. 47, 48

[182] Fava, Felipe Bedinotto, Luiz Felipe Laviola Leite, Luís Fernando Alves Da Silva, Pedro Ramires Da Silva Amalfi Costa, Angelo Gaspar Diniz Nogueira, Amanda Fagundes Gobus Lopes, Claudio Schepke, Diego Luis Kreutz, and Rodrigo Brãndao Mansilha: *Assessing the Performance of Docker in Docker Containers for Microservice-Based Architectures*. In *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pages 137–142, March 2024. https://ieeexplore.ieee.org/document/10495554, visited on 2025-06-02, ISSN: 2377-5750. 47

[183] Python Software Foundation: *Python programming language*. https://www.python.org/, 2025. Accessed: 2025-10-28. 48

[184] Kerney, Jamison, Ioan Raicu, John Raicu, and Kyle Chard: *Towards Fine-Grained Parallelism in Parallel and Distributed Python Libraries*. In *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 706–715, May 2024. https://ieeexplore.ieee.org/abstract/document/10596422, visited on 2025-06-03. 48

[185] Apache Software Foundation: *Apache spark*. https://spark.apache.org/, 2025. Accessed: 2025-10-28. 48

[186] Prometheus Authors: *Prometheus: Monitoring system & time series database*. https://prometheus.io/, 2025. Accessed: 2025-10-28. 48

[187] Grafana Labs: *Grafana: The open observability platform*. https://grafana.com/, 2025. Accessed: 2025-10-28. 48

[188] Dask Development Team: *Dask: Scalable analytics in python*. https://www.dask.org/, 2025. Accessed: 2025-10-28. 48

[189] Apache Software Foundation: *Apache parquet: Columnar storage for the hadoop ecosystem*. https://parquet.apache.org/, 2025. Accessed: 2025-10-28. 48

[190] Docker Inc.: *Docker: Empowering app development for developers*. https://www.docker.com/, 2025. Accessed: 2025-10-28. 48

[191] Crist, James: *Dask & Numba: Simple libraries for optimizing scientific python code*. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2342–2343, December 2016. https://ieeexplore.ieee.org/document/7840867, visited on 2025-06-03. 48

[192] Kumar, Mandeep: *Distributed Execution of Dask on HPC: A Case Study*. In *2023 World Conference on Communication & Computing (WCONF)*, pages 1–4, July 2023. https://ieeexplore.ieee.org/document/10234994, visited on 2025-05-04. 48

[193] Rezanejad, Amin and Amir Seyed Danesh: *Improving the Performance of the K-Nearest Neighbors Algorithm with Parallelization in Dask*. In *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5, February 2024. https://ieeexplore.ieee.org/document/10475304, visited on 2025-05-04, ISSN: 2640-5768. 48

[194] RabbitMQ Team: *Rabbitmq: Open source message broker*. https://www.rabbitmq.com/, 2025. Accessed: 2025-10-28. 48

[195] Maharjan, Rokin, Md Showkat Hossain Chy, Muhammad Ashfakur Arju, and Tomas Cerny: *Benchmarking Message Queues*. Telecom, 4(2):298–312, June 2023, ISSN 2673-4001. https://www.mdpi.com/2673-4001/4/2/18, visited on 2025-06-05, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 48

[196] Shi, Lukui, Jun Zhang, Enhai Liu, and Pilian He: *Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines*. In *Third International Conference on Natural Computation (ICNC 2007)*, volume 1, pages 674–677, August 2007. https://ieeexplore.ieee.org/document/4344276, visited on 2025-06-03, ISSN: 2157-9563. 48

[197] Swarnalatha, K, N Vinay Kumar, D S Guru, and B S Anami: *Analysis of Dimensionality Reduction Techniques for Effective Text Classification*. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–5, June 2021. https://ieeexplore.ieee.org/document/9498287, visited on 2025-06-03. 48

[198] Atzberger, Daniel, Tim Cech, Matthias Trapp, Rico Richter, Willy Scheibel, Jürgen Döllner, and Tobias Schreck: *Large-Scale Evaluation of Topic Models and Dimensionality Reduction Methods for 2D Text Spatialization*. IEEE Transactions on Visualization and Computer Graphics, 30(1):902–912, January 2024, ISSN 1941-0506. https://ieeexplore.ieee.org/document/10290971, visited on 2025-06-03. 48

[199] Boyapati, Mallika and Ramazan Aygun: *Semanformer: Semantics-aware Embedding Dimensionality Reduction Using Transformer-Based Models*. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 134–141, February 2024. https://ieeexplore.ieee.org/document/10475663, visited on 2025-06-03, ISSN: 2472-9671. 48

[200] Wei, Yuanzhen: *Research on Text Classification Based on Word Vector Models*. In *2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, volume 9, pages 489–493, November 2024. https://ieeexplore.ieee.org/document/10792865, visited on 2025-06-03, ISSN: 2189-8723. 48

[201] Modi, Anshul, Yuvraj Singh Dhanjal, and Anamika Larhgotra: *Semantic Similarity for Text Comparison between Textual Documents or Sentences*. In *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5, December 2023. https://ieeexplore.ieee.org/document/10465440, visited on 2025-06-04. 48

[202] Radeva, Irina, Ivan Popchev, Lyubka Doukovska, and Miroslava Dimitrova: *Web Application for Retrieval-Augmented Generation: Implementation and Testing*. Electronics, 13(7):1361, January 2024, ISSN 2079-9292. https://www.mdpi.com/2079-9292/13/7/1361, visited on 2025-06-08, Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. 49

[203] Corrêa, Nicholas Kluge, Aniket Sen, Sophia Falk, and Shiza Fatimah: *Tucano: Advancing Neural Text Generation for Portuguese*, November 2024. http://arxiv.org/abs/2411.07854, visited on 2025-06-04, arXiv:2411.07854 [cs]. 49

[204] Öztürk, Emir and Altan Mesut: *PERFORMANCE ANALYSIS OF CHROMA, QDRANT, AND FAISS DATABASES | Request PDF*. In *Proceedings of the UNITECH International Scientific Conference*, April 2024. https://www.researchgate.net/publication/387206678_PERFORMANCE_ANALYSIS_OF_CHROMA_QDRANT_AND_FAISS_DATABASES. 52

[205] Qdrant: *Qdrant - vector search engine for the next generation of ai*. https://qdrant.tech/, 2025. Accessed: 2025-06-08. 52

[206] Belani, Hrvoje, Marin Vukovic, and Zeljka Car: *Requirements Engineering Challenges in Building AI-Based Complex Systems*. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 252–255, September 2019. https://ieeexplore.ieee.org/abstract/document/8933653, visited on 2026-03-29. 60

[207] Shahriar, Sakib, Sonal Allana, Seyed Mehdi Hazratifard, and Rozita Dara: *A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle*. IEEE Access, 11:61829–61854, 2023, ISSN 2169-3536. https://ieeexplore.ieee.org/abstract/document/10155147, visited on 2026-03-29. 60, 61, 62, 71, 74

[208] Massey, Aaron K., Jacob Eisenstein, Annie I. Antón, and Peter P. Swire: *Automated text mining for requirements analysis of policy documents*. In *2013 21st IEEE International Requirements Engineering Conference (RE)*, pages 4–13, July 2013. https://ieeexplore.ieee.org/document/6636700, visited on 2026-03-11, ISSN: 2332-6441. 60

[209] Qiang, Jipeng, Ping Chen, Wei Ding, Tong Wang, Fei Xie, and Xindong Wu: *Topic Discovery from Heterogeneous Texts*. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 196–203, November 2016. https://ieeexplore.ieee.org/document/7814599, visited on 2026-03-08, ISSN: 2375-0197. 60

[210] Blei, David M., Andrew Y. Ng, and Michael I. Jordan: *Latent dirichlet allocation*. J. Mach. Learn. Res., 3(null):993–1022, March 2003, ISSN 1532-4435. https://dl.acm.org/doi/10.5555/944919.944937, visited on 2026-03-08. 61, 62, 71, 72, 74

[211] Kumar, Niraj, R.R Suman, and Sanjay Kumar: *Text Classification and Topic Modelling of Web Extracted Data*. In *Glob. Conf. Adv. Technol., GCAT*. Institute of Electrical and Electronics Engineers Inc., 2021, ISBN 978-0-7381-3215-0. Journal Abbreviation: Glob. Conf. Adv. Technol., GCAT. 61

[212] Vatsalan, Dinusha, Raghav Bhaskar, Aris Gkoulalas-Divanis, and Dimitrios Karapiperis: *Privacy Preserving Text Data Encoding and Topic Modelling*. In *Proc. - IEEE Int. Conf. Big Data, Big Data*, pages 1308–1316. Institute of Electrical and Electronics Engineers Inc., 2021, ISBN 978-1-6654-3902-2. Journal Abbreviation: Proc. - IEEE Int. Conf. Big Data, Big Data. 61

[213] Blei, David M.: *Probabilistic topic models*. Commun. ACM, 55(4):77–84, April 2012, ISSN 0001-0782. https://dl.acm.org/doi/10.1145/2133806.2133826, visited on 2026-03-08. 61, 62

[214] Astudillo, Gabriel: *A Large Language Model Approach for In-Depth Qualitative Text Analysis*. In *2025 15th IEEE International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7, December 2025. https://ieeexplore.ieee.org/document/11302863, visited on 2026-03-27. 61

[215] Lewis, Craig M. and Francesco Grossetti: *A statistical approach for optimal topic model identification*. J. Mach. Learn. Res., 23(1):58:2553–58:2572, January 2022, ISSN 1532-4435. https://dl.acm.org/doi/10.5555/3586589.3586647, visited on 2026-03-08. 61

[216] Griffiths, Thomas L. and Mark Steyvers: *Finding scientific topics*. Proceedings of the National Academy of Sciences, 101(suppl_1):5228–5235, April 2004. https://www.pnas.org/doi/abs/10.1073/pnas.0307752101, visited on 2026-03-10. 61

[217] Bagul, Dhiraj Vaibhav and Sunita Barve: *A novel content-based recommendation approach based on LDA topic modeling for literature recommendation*. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 954–961, January 2021. https://ieeexplore.ieee.org/document/9358561, visited on 2026-03-13. 62

[218] Goyal, Astha and Indu Kashyap: *A data-driven analysis to determine the optimal number of topics 'K' for latent Dirichlet allocation model*. Indonesian Journal of Electrical Engineering and Computer Science, 35(1):310, July 2024, ISSN 2502-4760, 2502-4752. https://ijeecs.iaescore.com/index.php/IJEECS/article/view/35624, visited on 2026-03-27. 62, 66, 71

[219] Qiang, Ji Peng, Ping Chen, Wei Ding, Fei Xie, and Xindong Wu: *Multi-document summarization using closed patterns*. Knowledge-Based Systems, 99:28–38, May 2016, ISSN 0950-7051. https://www.sciencedirect.com/science/article/pii/S0950705116000502, visited on 2026-03-10. 62

[220] Wang, Qiaozhi, Hao Xue, Fengjun Li, Dongwon Lee, and Bo Luo: *#DontTweetThis: Scoring Private Information in Social Networks*. Proceedings on Privacy Enhancing Technologies, 2019, ISSN 2299-0984. https://petsymposium.org/popets/2019/popets-2019-0059.php, visited on 2026-03-26. 62

[221] Mao, Huina, Xin Shuai, and Apu Kapadia: *Loose tweets: an analysis of privacy leaks on twitter*. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, WPES '11, pages 1–12, New York, NY, USA, October 2011. Association for Computing Machinery, ISBN 978-1-4503-1002-4. https://dl.acm.org/doi/10.1145/2046556.2046558, visited on 2026-03-26. 62

[222] Tillmann, Arne, Lindrit Kqiku, Delphine Reinhardt, Christoph Weisser, Benjamin Säfken, and Thomas Kneib: *Privacy Estimation on Twitter: Modelling the Effect of Latent Topics on Privacy by Integrating XGBoost, Topic and Generalized Additive Mod-*

els. In *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, pages 2325–2332, December 2022. https://ieeexplore.ieee.org/document/10189773, visited on 2026-03-25. 62

[223] Alemany, Jose, Elena Del Val, and Ana García-Fornes: *Empowering users regarding the sensitivity of their data in social networks through nudge mechanisms*. Hawaii International Conference on System Sciences 2020 (HICSS-53), January 2020. https://aisel.aisnet.org/hicss-53/dsm/decision_making_in_osn/3. 62

[224] Hiniduma, Kaveen, Suren Byna, and Jean Luca Bez: *Data Readiness for AI: A 360-Degree Survey*. ACM Comput. Surv., 57(9):219:1–219:39, April 2025, ISSN 0360-0300. https://dl.acm.org/doi/10.1145/3722214, visited on 2026-03-30. 63

[225] Martinelli, Fabio, Fiammetta Marulli, Francesco Mercaldo, Stefano Marrone, and Antonella Santone: *Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence*. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2020. 63

[226] OpenAI: *Chatgpt (gpt-5.4 thinking)*, 2026. https://chatgpt.com/, Large language model, accessed 2026-03-26. 65

[227] Vukanti, Vidya and Anu Jose: *Business Analytics: A case-study approach using LDA topic modelling*. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1818–1823, April 2021. https://ieeexplore.ieee.org/document/9418344, visited on 2026-03-26. 65

[228] Explosion AI: *spacy*. https://spacy.io/, n.d. Official website. Accessed: 2026-03-26. 65

[229] Google: *Anti-gravity*. https://antigravity.google/. Accessed: 2026-03-23. 66

[230] Anthropic: *Claude sonnet 4.6*. https://www.anthropic.com/news/claude-sonnet-4-6. Anthropic News. Accessed: 2026-03-23. 66

[231] Google DeepMind: *Gemini pro*. https://deepmind.google/models/gemini/pro/. Accessed: 2026-03-23. 66

[232] PyPI: *gensim*. https://pypi.org/project/gensim/, visited on 2026-03-13, Python Package Index project page. 66

[233] Řehůřek, Radim: *gensim*. https://github.com/piskvorky/gensim, n.d. GitHub repository. Accessed: 2026-03-13. 66

[234] Řehůřek, Radim and Petr Sojka: *Software Framework for Topic Modelling with Large Corpora*. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en. 66

[235] Řehůřek, Radim: *Gensim: Topic modelling for humans*. https://radimrehurek.com/gensim/, n.d. Official documentation website. Accessed: 2026-03-13. 66

[236] Ollama: *llama3.1*. https://ollama.com/library/llama3.1, visited on 2026-03-26, Ollama model library page. 68

[237] Ollama: *Ollama*. https://ollama.com/, visited on 2026-03-26, Official website. 68

[238] Egger, Roman and Joanne Yu: *A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts*. Frontiers in Sociology, 7, May 2022, ISSN 2297-7775. https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full, visited on 2026-03-21. 70

[239] Hossain, Md. and Douglas Timmer: *Machine Learning Model Optimization with Hyper Parameter Tuning Approach*. Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence, January 2021. https://scholarworks.utrgv.edu/mie_fac/107. 71

[240] Ianina, Anastasia and Konstantin Vorontsov: *Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search*. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 131–138, November 2019. https://ieeexplore.ieee.org/abstract/document/8981493, visited on 2026-03-27, ISSN: 2305-7254. 71