Privacy Preservation in Textual Data: A Systematic Mapping Study on Differential Privacy and Semantic Similarity

Daniel Linhares Lim-Apo, Edna Dias Canedo

[1]University of Brasília (UnB), Department of Computer Science, Brasília, DF, Brazil
E-mail: daniel.unb@sede.com.br, ednacanedo@unb.br

Abstract. ***Background:*** *AI and machine learning increasingly depend on large-scale textual data, which often embeds personally identifiable and sensitive information. In this context, privacy-preserving processing of unstructured text has become essential to mitigate disclosure and re-identification risks, especially in pipelines that rely on semantic representations and similarity measures.* ***Goal:*** *This study aims to map state-of-the-art techniques for privacy-preserving textual data analysis and to characterize how privacy mechanisms relate to semantic similarity and the identification of infrequent (rare) textual patterns.* ***Method:*** *We conducted a Systematic Mapping Study (SMS) following established guidelines, retrieving peer-reviewed publications from four major digital libraries (2010–2025). The selected studies were screened using inclusion/exclusion criteria and a quality checklist, and the extracted data were synthesized through frequency-based and thematic mapping aligned with three research questions: (i) privacy-preserving techniques used in textual data analysis, (ii) computational approaches (data science and language-model-based methods) supporting such mechanisms, and (iii) techniques adopted for semantic similarity and rare-event oriented text analysis under privacy constraints.* ***Results:*** *The mapping shows that anonymization/de-identification, differential privacy, and federated learning are among the most recurrent privacy-preserving approaches reported for text. It also highlights the prevalence of NLP pipelines and transformer-based models (e.g., BERT variants and large language models) as supporting components, typically combined with classic semantic similarity techniques such as vector-space representations, embeddings, topic modeling, and cosine similarity. Rare-event detection appears less frequently, suggesting an emerging gap and opportunities for future research on privacy-aware pipelines for low-frequency phenomena in text. The results provide an evidence-based overview to support the design of guidelines and best practices for privacy-preserving text analytics in software intensive settings.*

1. Introduction

The increasing reliance on Artificial Intelligence (AI) and Machine Learning (ML) systems has intensified the need to process large volumes of data, much of which is expressed in textual form. While text enables rich semantic analysis, it also poses substantial privacy risks, as unstructured language often embeds explicit and implicit personal identifiers. Research on privacy-preserving techniques has traditionally addressed both structured and unstructured data [Aghasian et al. 2020], yet applying these techniques to textual data remains particularly challenging due to its high dimensionality and contextual sensitivity.

In text-based scenarios, privacy threats are closely associated with disclosure and re-identification risks. Prior work has shown that traditional privacy models, such as k-anonymity, are often insufficient in high-dimensional or unstructured settings [SWEENEY 2012, Lee et al. 2025]. Differential privacy has emerged as a prominent alternative, offering formal guarantees against information leakage [Dwork 2006, Cui et al. 2024]. However, enforcing such guarantees in textual data is non-trivial, as utility is strongly influenced by text frequency, semantic similarity, and the presence of rare or distinctive events.

Recent advances in natural language processing (NLP) and artificial intelligence (AI) have further reshaped this landscape. Modern text analytics increasingly rely on deep neural architectures, including transformer-based models and Large Language Models (LLMs) [Souza et al. 2023, Zhao et al. 2025]. At the same time, emerging paradigms such as agentic and multi-agent AI systems [Acharya et al. 2025, Duan and Wang 2024], together with supporting infrastructures like the Model Context Protocol (MCP) [Anthropic 2024, Khoei et al. 2025], have expanded the scale and autonomy of text processing pipelines. In parallel, significant research efforts have addressed anonymization and de-identification in text [Lison et al. 2021, Giampaolo et al. 2023, Murin et al. 2024, Asimopoulos et al. 2024], as well as techniques for detecting rare or infrequent events in textual data [Shyalika et al. 2024, Abubakar et al. 2024]. Despite this progress, these research streams remain fragmented, particularly regarding how privacy guarantees interact with semantic similarity and rarity analysis.

From a conceptual perspective, privacy is commonly defined as an individuals right to control self-disclosure and the extent to which personal information is revealed [Federative Republic of Brazil 2025]. This notion underpins core data protection principles, such as data minimization, anonymity, and informed consent [Kluge Corrêa 2024]. Nevertheless, many benefits of data-driven technologies depend on collecting and analyzing sensitive information [Mendes and Vilela 2017]. In textual datasets, this tension is amplified, as language frequently conveys contextual and rare attributes that can expose individuals even after partial sanitization. Protecting sensitive content in unstructured text before it is shared with untrusted parties therefore remains a central challenge in contemporary privacy research [Zhao and Chen 2022].

Data mining aims to discover patterns and construct models from observed data [Fayyad et al. 1996]. Privacy-preserving techniques inevitably alter the original data, which may weaken the correspondence between learned models and the underlying phenomena. In statistical learning, this trade-off between privacy and utility has become increasingly critical, particularly when large-scale confidential data are processed or shared under regulatory constraints [Zhou et al. 2009]. Stronger privacy guarantees often reduce analytical value, highlighting the need to better understand how existing techniques balance confidentiality and utility in textual data analysis.

Against this background, this paper presents a Systematic Mapping Study that synthesizes existing research on privacy-preserving techniques for textual data analysis, with particular attention to semantic similarity and rare-event-oriented methods under privacy constraints.

2. Systematic Mapping Study (SMS)

We conducted a Systematic Mapping Study (SMS) following the protocol proposed by Kitchenham and Charters [Barbara and Charters 2007], with the objective of identifying and organizing prevailing techniques for privacy-preserving analysis of textual data. The SMS methodology provides a structured, transparent, and replicable process for systematically identifying, classifying, and synthesizing existing research within a defined domain. According to Kitchenham and Charters [Barbara and Charters 2007], the SMS process is organized into three main phases: *Planning*, *Conducting*, and *Reporting*.

The SMS was conducted with the objective of identifying contemporary and state-of-the-art practices in privacy-preserving techniques applied to textual data [Mendes and Vilela 2017]. The review is guided by the overarching problem of how unstructured textual data can be protected when used in artificial intelligence models, while preserving analytical utility and ensuring the privacy of individuals. Specifically, the study examines how personal data in text can be safeguarded through the adoption of privacy-preserving techniques, data science methods (e.g., vectorization and representation learning), and AI-based approaches, including LLMs and agent-based models. In addition, the review investigates which techniques are employed for semantic similarity analysis and rare-event detection in textual data under differential privacy constraints. Identifying the privacy-preserving techniques that have been applied in text analysis is therefore central to addressing this research problem.

To operationalize this objective, it was decomposed into three interrelated research questions (RQ):

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.

Search String Petticrew and Roberts [Petticrew and Roberts 2008] propose that research questions should be structured using five elements, commonly referred to as PICOC: Population, Intervention, Comparison, Outcome, and Context. This framework is widely adopted in secondary studies to support the formulation of research questions and the systematic construction of search strategies.

As described in Table 1. PICOC at Zenodo, the Population was defined as textual datasets containing facts or events with sensitive content. The Intervention element

encompasses privacy-preserving techniques combined with data science methods, large language models, and agent-based AI approaches applied to semantic similarity analysis and rarity detection in textual data. The Comparison element of the PICOC framework was not specified for this study, as the objective of the SMS is to survey and synthesize existing approaches rather than to benchmark alternative techniques. Consequently, no explicit comparison criteria were defined, as documented in Table 1. PICOC.

The Outcome element targets improvements in privacy-aware textual data analysis while preserving analytical utility. As summarized in Table 1. PICOC, the expected outcomes include protection of sensitive information, preservation of semantic richness, identification of rare or distinctive textual patterns, and an improved balance between privacy guarantees and data usability. The Context element focuses on sensitive or high-risk application domains in which textual data analysis requires strong privacy safeguards. These contexts include AI-driven natural language processing in regulated environments, legal and compliance analytics, and scenarios involving re-identification risks under differential privacy principles, as detailed in Table 1. PICOC.

Based on the defined PICOC elements, a search string was constructed by combining Population, Intervention, Outcome, and Context components using Boolean operators. This strategy enabled systematic querying across multiple digital libraries to capture studies addressing privacy-preserving text analysis, semantic similarity, privacy utility trade-offs, and applications in regulated or sensitive domains.

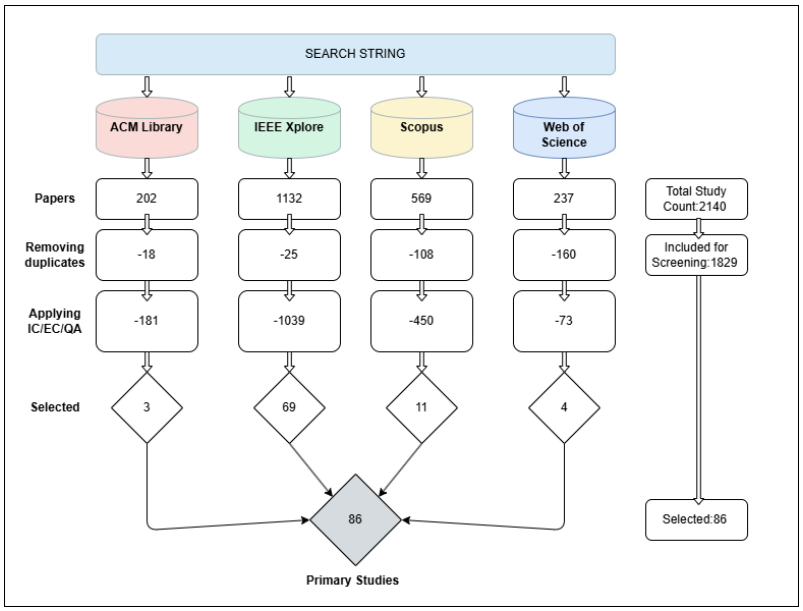
The search string was iteratively refined through exploratory searches to assess result relevance and identify additional keywords. The final generic search string, together with database-specific adaptations, is openly available in Zenodo at Search-String. The digital libraries selected to execute the search string were the ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. The decision to use these four databases was guided by recommendations from Kitchenham et al. [Brereton et al. 2007], who identify them as core sources for secondary studies in Software Engineering. In addition, Merrouni et al. [Merrouni et al. 2016] highlight that ACM, IEEE, and Scopus host a substantial portion of high-quality research in computer science. The search string was adapted to the syntax of each digital library, and access to all databases was provided through the CAPES Periodicals Portal using institutional credentials.

To select relevant studies, we applied predefined Inclusion Criteria (IC), whereby studies were included if they satisfied the conditions described in Table 2. Inclusion Criteria. Similarly, Exclusion Criteria (EC) were applied to remove studies that met any of the conditions listed in Table 3. Exclusion Criteria, both openly available in Zenodo. While the inclusion and exclusion criteria ensured topical relevance, an additional quality assessment was conducted to verify the methodological soundness and evaluative rigor of the selected studies. For this purpose, we adopted a quality assessment checklist, with evaluation criteria defined in Table 4. Study Quality Assessment Criteria.

Each study was assessed using a simple scoring rubric for each quality criterion: **High** (3) when the criterion was clearly and comprehensively addressed; **Medium** (2) when it was partially addressed with limitations; **Low** (1) when it was weakly addressed; and **Absent** (0) when it was not addressed. A study was excluded if it failed to achieve at least a High score in QAC-1, which evaluates relevance to privacy, semantic similarity,

or rarity dimensions, or at least a Medium score in QAC-2, which assesses the clarity and rigor of the research methodology. In addition, a minimum overall cutoff score of 2 was required for inclusion in the final set of studies.

Conducting To manage and operationalize the Systematic Mapping Study, we used Parsifal [Parsifal Developers 2025], a free and open-source web platform specifically designed to support secondary studies in software engineering. Parsifal is aligned with the methodological guidelines proposed by Kitchenham and Charters [Barbara and Charters 2007] and supports the main activities of an SMS, including study screening, collaborative reviewing, and duplicate detection. In particular, the platform facilitates efficient navigation through titles and abstracts during the selection phase and ensures traceability of decisions throughout the review process. Figure 1 presents the number of studies retained at each stage of the mapping.



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

Figure 1. Remaining papers after each step of the SMS.

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.

Data Extraction: The data extraction stage in a SMS refers to the process 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 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 extraction aimed to capture key dimensions of the object of study. 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 5. Data Extraction Form available at Zenodo. The synthesis of the included studies followed a mixed-methods approach comprising both qualitative and quantitative elements: narrative synthesis, tabular comparison, thematic mapping and meta analysis.

3. Results and Discussion

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 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 research questions. This framework encompasses the principal techniques, methodologies, processes, theoretical models, and tools identified across the corpus of studies.

3.1. 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 6. Privacy-preserving Techniques (RQ.1) in Zenodo

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, as shown in Figure 2. Other relevant ones include Differential Privacy, Encryption, and K-Anonymity, as shown in Table 6. Privacy-preserving Techniques (RQ.1) in Zenodo.

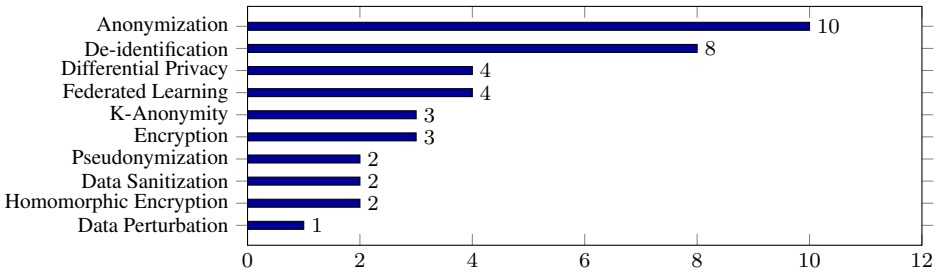


Figure 2. Top 10 most frequently cited privacy-preserving techniques in the SMS (RQ.1).

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 [Union 2025] and LGPD [Federative Republic of Brazil 2025], these concepts differ.

The use of textual data is relevant in numerous domains, and data sharing is essential. However, it raises serious privacy concerns when the data contain personal information [Hassan et al. 2019]. Data anonymization aims to preserve privacy while allowing data dissemination preventing individual identification and maintains analytical utility [Giampaolo et al. 2023]. Although several techniques have been developed for anonymizing structured data, automatic anonymization of unstructured textual data remains unresolved and far from being fully achieved [Hassan et al. 2019]. Privacy models 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 [Lison et al. 2021]. Text anonymization typically involves identifying sensitive elements that are removed or generalized to protect individual privacy [Hassan et al. 2019].

Asimopoulos et al. [Asimopoulos et al. 2024] 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 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 [Volodina et al. 2023]. The GDPR explicitly identifies pseudonymization for obscuring an individuals identity in data processing and recognizes pseudonymization as a practical measure to demonstrate compliance with key obligations, such as data protection by design [Volodina et al. 2023].

Federated Learning (FL) [Khan et al. 2024, Liu et al. 2020] represents a distributed machine learning paradigm designed to enable model training while mitigating privacy risks. FL [Asif et al. 2024, Lim et al. 2020, Tan et al. 2023] allows multiple edge users to train a global model without exchanging raw data, thereby preserving user privacy. The training process involves iterative local model updates and global aggregation.

In this context, Liu et al. [Liu et al. 2020] explain that traditional centralized learning methods typically involve three sequential stages: data preprocessing, data integration, and model construction. The process usually includes sample selection, outlier elimination, feature normalization, and feature combination. The subsequent phase involves directly sharing datasets among entities to generate a unified global dataset for training. This centralized approach poses significant challenges under modern data protection frameworks, as the exchange of raw data across organizations can disclose sensitive information and may violate privacy regulations such as the GDPR. Privacy-preserving techniques identified in the SMS for textual data analysis (RQ.1) are classified in clusters in the file Table 6. Privacy-preserving Techniques (RQ.1) openly available in Zenodo.

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.

3.2. 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 7. Privacy-preserving Techniques (RQ.2) openly available in Zenodo 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 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), as shown in Figure 3.

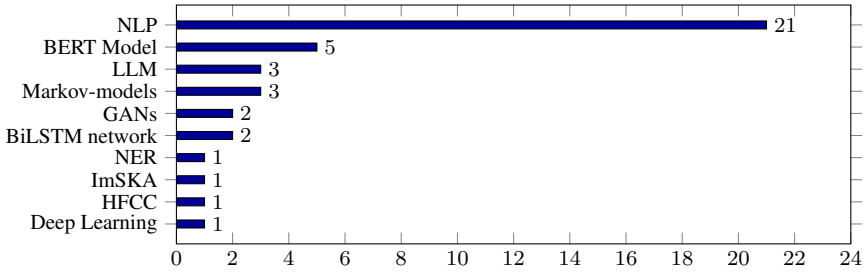


Figure 3. Top 10 most frequently cited computational approaches in the SMS (RQ.2).

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 [Languré and Zareei 2025].

LLM-oriented pipelines offer another avenue. Gupta et al. [Gupta et al. 2024] 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 7. Privacy-preserving Techniques (RQ.2) openly available in Zenodo.

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 privacyutility trade-off. Cryptographic preprocessing before LLM interaction strengthens protection of identifiers without materially degrading downstream tasks.

3.3. 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 8. Privacy-preserving Techniques (RQ.3) in Zenodo 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*, *TFIDF*, *Support Vector Machines (SVM)*, *Part-of-Speech (POS) Tagging*, *Graph Creation*, *Short Text Understanding*,*GloVe*, and *Word2Vec*, as shown in Figure 4.

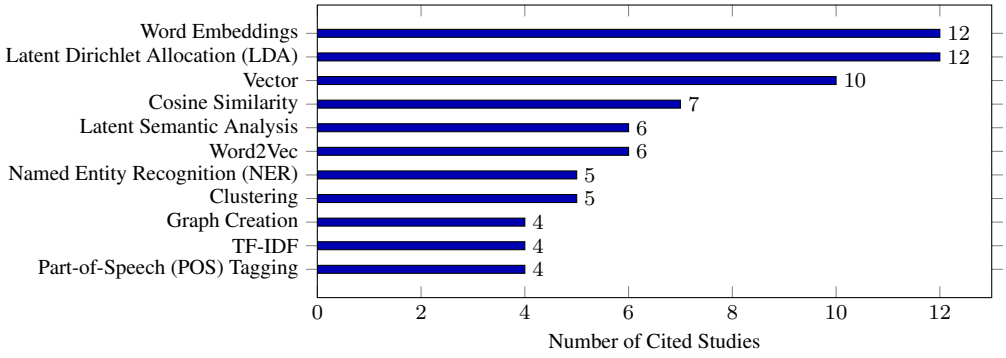


Figure 4. Top 11 most frequently cited semantic comparison techniques in the SMS (RQ.3).

Table 6. Privacy-preserving Techniques (RQ.1) openly available in Zenodo shows that 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 9. Innovative Techniques in Zenodo, as they may be seeds for future research and work. 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 TFIDF, 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. Integrating DP into semantic similarity is feasible with modest utility loss; extending it to rare-event detection requires tailored data and evaluation protocols.

3.4. 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 threats to validity and the measures adopted to mitigate them, in line with the methodological standards of literature reviews [Barbara and Charters 2007]. **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 syn-

thesis 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.

4. Conclusion and Future Work

This study presented a Systematic Mapping Study that synthesized and organized the current body of knowledge on privacy-preserving techniques for textual data analysis. The results show that a broad range of approaches has been proposed to mitigate privacy risks in text-based AI pipelines, with particular emphasis on anonymization and de-identification strategies, differential privacy mechanisms, and federated learning. The mapping further reveals that most solutions rely on established NLP and semantic similarity techniques, such as vector-space representations, embeddings, and topic modeling, while explicitly privacy-aware support for rare-event detection in text remains comparatively underexplored.

By consolidating evidence across primary studies, this review provides a structured overview of how privacy-preserving mechanisms are currently applied to unstructured textual data, highlighting both dominant practices and emerging gaps. In particular, the imbalance between well-established semantic analysis techniques and the limited attention given to rare-event detection under privacy constraints points to important opportunities for future research in sensitive and regulated domains.

As future work, we plan to derive a structured framework of guidelines to support the selection and application of privacy-preserving techniques throughout the textual data lifecycle, including data collection, processing, storage, and dissemination. In addition, systematic comparative studies are needed to empirically assess trade-offs between privacy guarantees and analytical utility, as well as the suitability of different techniques across application contexts and regulatory environments.

Acknowledgements

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant N.300883/2025-0 and 406266/2025-5.

Artifact Availability

All research materials supporting the findings of this Systematic Mapping Study, including the study protocol, screening records, and extracted datasets, are openly available on Zenodo at https://Zenodo.org/records/18498099

References

Abubakar, Y. I., Othmani, A., Siarry, P., and Sabri, A. Q. M. (2024). A Systematic Review of Rare Events Detection Across Modalities Using Machine Learning and Deep Learning. *IEEE Access*, 12:47091–47109.

Acharya, D. B., Kuppan, K., and Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex GoalsA Comprehensive Survey. *IEEE Access*, 13:18912–18936.

Aghasian, E., Garg, S., and Montgomery, J. (2020). An automated model to score the privacy of unstructured informationSocial media case. *Computers & Security*, 92:101778.

Anthropic (2024). Model context protocol. `https://modelcontextprotocol.io/introduction`. Accessed: April 9, 2025.

Asif, H., Min, S., Wang, X., and Vaidya, J. (2024). U.S.-U.K. PETs Prize Challenge: Anomaly Detection via Privacy-Enhanced Federated Learning. *IEEE Transactions on Privacy*, 1:3–18. Conference Name: IEEE Transactions on Privacy.

Asimopoulos, D., Siniosoglou, I., Argyriou, V., Karamitsou, T., Fountoukidis, E., Goudos, S. K., Moscholios, I. D., Psannis, K. E., and Sarigiannidis, P. (2024). 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. ISSN: 2993-4443.

Barbara, K. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Keele University, UK*, 9:1–65.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(4):571–583.

Cui, J., Shen, H., and Cao, Y. (2024). Survey on the Applications of Differential Privacy. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 43–47.

Duan, Z. and Wang, J. (2024). Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI. arXiv:2411.18241 [cs].

Dwork, C. (2006). Differential Privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Federative Republic of Brazil (2025). Lei Geral de Protecao de Dados Pessoais (LGPD) - Lei nž 13.709, de 14 de agosto de 2018. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm`. Accessed: 2025-03-27.

Giampaolo, F., Izzo, S., Prezioso, E., Chiaro, D., Cuomo, S., Bellandi, V., and Piccialli, F. (2023). 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. ISSN: 2837-0740.

Gupta, B. B., Gaurav, A., Arya, V., Alhalabi, W., Alsalman, D., and Vijayakumar, P. (2024). Enhancing user prompt confidentiality in Large Language Models through advanced differential encryption. *Computers and Electrical Engineering*, 116:109215.

Hassan, F., Sánchez, D., Soria-Comas, J., and Domingo-Ferrer, J. (2019). 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. ISSN: 2324-9013.

Khan, Y., Sánchez, D., and Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: a systematic literature review. *Artificial Intelligence Review*, 57(12):320.

Khoei, T. T., Ehtesham, A., Kumar, S., and Khoei, T. T. (2025). A Survey of the Model Context Protocol (MCP): Standardizing Context to Enhance Large Language Models (LLMs).

Kluge Corrêa, N. (2024). *Dynamic Normativity*. Thesis, Universitäts- und Landesbibliothek Bonn. Accepted: 2024-06-11T12:54:16Z.

Languré, A. d. L. and Zareei, M. (2025). Privacy-Preserving Emotion Detection: Evaluating the Trade-Off Between K-Anonymity and Model Performance. *IEEE Access*, 13:105901–105910.

Lee, S., Kim, Y., Kwon, Y., and Cho, S. (2025). Secure privacy-preserving record linkage system from re-identification //attack. *PLOS ONE*, 20(1):e0314486. Publisher: Public Library of Science.

Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., and Miao, C. (2020). Federated Learning in Mobile Edge Networks: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063. Conference Name: IEEE Communications Surveys & Tutorials.

Lison, P., Pilán, I., Sanchez, D., Batet, M., and Øvrelid, L. (2021). Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *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. Association for Computational Linguistics.

Liu, Y., Yu, J. J. Q., Kang, J., Niyato, D., and Zhang, S. (2020). Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach. *IEEE Internet of Things Journal*, 7(8):7751–7763.

Mendes, R. and Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access*, 5:10562–10582. Conference Name: IEEE Access.

Merrouni, Z. A., Frikh, B., and Ouhbi, B. (2016). 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. ISSN: 2327-1884.

Murin, M., Molan, S., Michalkc, M., Kainz, O., and Cymbalák, D. (2024). 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.

Parsifal Developers (2025). Parsifal: A platform for formal modeling and verification of privacy-preserving systems. `https://parsif.al`. Accessed: 2025-09-21.

Petticrew, M. and Roberts, H. (2008). *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.

Shyalika, C., Wickramarachchi, R., and Sheth, A. (2024). A Comprehensive Survey on Rare Event Prediction. arXiv:2309.11356 [cs].

Souza, F. C., Nogueira, R. F., and Lotufo, R. A. (2023). BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901.

SWEENEY, L. (2012). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. Publisher: World Scientific Publishing Company.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. (2023). Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Union, E. (2025). General data protection regulation - gdpr. `https://gdpr-info.eu/`. Accessed: 2025-03-27.

Volodina, E., Dobnik, S., Tiedemann, T. L. m., and Vu, X.-S. (2023). 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.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2025). A Survey of Large Language Models. arXiv:2303.18223 [cs].

Zhao, Y. and Chen, J. (2022). A survey on differential privacy for unstructured data content. *ACM Comput. Surv.*, 54(10s):207:1–207:28.

Zhou, S., Ligett, K., and Wasserman, L. (2009). Differential privacy with compression. In *2009 IEEE International Symposium on Information Theory*, pages 2718–2722. ISSN: 2157-8117.