Privacy-Preserving Semantic Document Retrieval: A Survey on SBERT, Federated Learning, and Homomorphic Encryption

Rinki Barman, Ruchi Chaturvedi
Computer Science & Engineering, Sam Global University, Bhopal, India

Abstract: The rapid growth of digital information has intensified the demand for secure and efficient document retrieval systems, particularly in domains such as healthcare, law, and finance, where data sensitivity is paramount. Traditional keyword-based methods like TF-IDF and BM25 provide effective baseline retrieval but fail to capture semantic meaning. Advances in deep learning, particularly through BERT and its extension SBERT, have enabled semantic embeddings that significantly improve contextual relevance. However, deploying such models in privacy-sensitive environments introduces new challenges. This survey provides a comprehensive overview of privacy-preserving semantic document retrieval, focusing on the intersection of SBERT-based embeddings, Federated Learning (FL), and Homomorphic Encryption (HE). We review the foundations of retrieval methods, explore encryption-based and federated frameworks for securing retrieval pipelines, and highlight the role of metaheuristic optimisation techniques (PSO, GA, ACO, BOA) in balancing accuracy, efficiency, and security. Finally, we identify key research gaps, including the trade-off between privacy and accuracy, clustering in encrypted space, and federated–encrypted integration. We discuss future directions for designing scalable, secure, and intelligent retrieval systems.

Keywords: Privacy-preserving retrieval, Semantic embeddings, SBERT, Federated learning, Homomorphic encryption, Searchable encryption

1. **Introduction**

The exponential growth of digital information has made efficient and accurate document retrieval an essential component of modern information systems. Search engines, digital libraries, enterprise knowledge bases, and intelligent assistants all depend on robust retrieval mechanisms to serve users with relevant information from vast document collections. Traditionally, retrieval relied heavily on keyword matching techniques, such as Term Frequency–Inverse Document Frequency (TF-IDF) and BM25. While these models achieved considerable success, they often struggled to capture semantic meaning, thereby failing to retrieve documents that were relevant but expressed in different lexical forms. For example, a query about “global warming” may overlook documents discussing “climate change,” despite their conceptual overlap. This limitation gave rise to semantic retrieval approaches, powered initially by distributed representations like Word2Vec and later by deep contextual models such as BERT and its variants. Among these, Sentence-BERT (SBERT) has emerged as a particularly powerful framework for encoding documents and queries into semantically meaningful vector spaces, enabling more precise and context-aware retrieval. Despite these advances, new challenges have surfaced, particularly around security and privacy. The sensitive nature of information in domains such as healthcare, law, finance, and defence underscores the importance of ensuring that retrieval processes do not expose private data. For instance, a medical research institute might need to perform semantic document retrieval across patient records or collaborative datasets without revealing raw data due to confidentiality and regulatory constraints such as GDPR or HIPAA. Similarly, enterprises may want to search across distributed corporate data repositories without risking data leakage or unauthorised access. This dual demand for high-quality semantic retrieval coupled with strong privacy guarantees has motivated a growing body of research into privacy-preserving retrieval mechanisms. The central problem lies in balancing the trade-off between retrieval effectiveness and data protection. While semantic models like SBERT excel in capturing deep contextual relationships, their deployment often requires access to raw documents for training and indexing. Encrypting or distributing data, on the other hand, introduces computational overhead and can degrade retrieval performance. Homomorphic Encryption (HE) promises to allow computations on encrypted data without decryption, but its efficiency remains a challenge for large-scale applications. Similarly, Federated Learning (FL) enables collaborative training across distributed nodes without sharing raw data, but ensuring secure aggregation and mitigating inference attacks are still open concerns. As such, researchers face the persistent tension between accuracy and privacy, where improving one often comes at the expense of the other. This survey article is motivated by the urgent need to consolidate and analyse these developments in privacy-preserving semantic document retrieval. While significant progress has been made in each of the individual strands, semantic representation learning, encryption schemes, and distributed training, there remains a lack of comprehensive synthesis connecting them. Existing surveys on document retrieval often focus either on traditional information retrieval or on semantic embedding techniques, without adequately addressing the privacy-preserving dimension. Conversely, studies in secure computation and encryption rarely explore their integration with advanced semantic retrieval methods. A holistic review is therefore necessary to map the intersection of these domains, highlight existing approaches, and expose gaps that require further innovation. The primary objectives of this article are threefold. First, it seeks to provide a systematic overview of the foundations of document retrieval, tracing the evolution from keyword-based approaches to deep semantic embeddings, with particular emphasis on SBERT and its multilingual and domain-adapted variants. Second, it examines the state of the art in privacy-preserving retrieval mechanisms, covering both cryptographic techniques such as Homomorphic Encryption and Searchable Encryption, and collaborative paradigms like Federated Learning. Third, it explores optimisation techniques, including metaheuristics such as Particle Swarm Optimisation (PSO) and Genetic Algorithms (GA) that can enhance retrieval performance under privacy constraints. Through this tripartite structure, the survey aims to equip researchers and practitioners with both conceptual understanding and practical insights into designing secure semantic retrieval systems. The novelty of this survey lies in its integrative perspective. Unlike prior work, it does not treat retrieval, encryption, and distributed learning as isolated research silos. Instead, it situates them within a unified framework that foregrounds their interdependencies and trade-offs. For instance, the article discusses how homomorphic operations can be tailored to vector similarity computations central to SBERT-based retrieval, or how federated architectures can be adapted for fine-tuning domain-specific embeddings without exposing sensitive data. Additionally, it highlights hybrid approaches, such as combining optimisation algorithms with privacy-preserving retrieval schemes, that have received limited attention but hold substantial potential. By weaving together these strands, the survey contributes a roadmap for advancing the field toward secure, efficient, and semantically rich retrieval systems. In summary, semantic document retrieval has evolved from simple keyword matching to sophisticated neural embeddings, but the imperative of privacy protection introduces fresh challenges and opportunities. This survey situates itself at this critical intersection, offering both a comprehensive review of existing methods and a forward-looking vision for future research.

2. Foundations of Document Retrieval

Document retrieval forms the backbone of modern information systems, enabling users to access relevant information from massive digital repositories. Over the decades, retrieval techniques have evolved from simple keyword matching to sophisticated semantic models that capture contextual meaning. This section reviews the two primary paradigms, keyword-based retrieval and semantic retrieval, outlining their theoretical underpinnings, strengths, and limitations.

1.1 Keyword-Based Retrieval

Term Frequency–Inverse Document Frequency(TF-IDF): One of the earliest and most widely adopted methods for document retrieval is TF-IDF. At its core, TF-IDF balances two key factors: the frequency of a term within a document (term frequency, TF) and the uniqueness of that term across the entire corpus (inverse document frequency, IDF). This weighting scheme emphasises words that occur frequently in a specific document but rarely across the corpus, thereby improving retrieval relevance. For example, in a medical database, the word diabetes may appear in many documents and thus be down-weighted. In contrast, gestational diabetes might occur in fewer documents and receive higher importance. The strengths of TF-IDF lie in its simplicity, efficiency, interpretability, and scalability for moderately sized corpora. However, it also has notable limitations, such as ignoring word order and syntactic relationships, failing to address synonyms and polysemous words, and showing reduced performance when applied to long, verbose queries or highly heterogeneous corpora. Despite these shortcomings, TF-IDF remains a foundational technique and continues to serve as a strong baseline in modern retrieval benchmarks. A Probabilistic Ranking Function (BM25): Building on TF-IDF, Okapi BM25 introduced a probabilistic framework to overcome limitations such as term frequency saturation and document length bias. Unlike TF-IDF, BM25 reduces the influence of terms that appear excessively in long documents, ensures that term relevance saturates rather than increasing linearly with frequency, and incorporates tunable parameters that allow retrieval behaviour to be adapted for different corpora. In practice, BM25 has proven to be a highly effective keyword-based retrieval method. It remains widely used in applications such as legal search, digital libraries, and as a baseline in modern search engines. However, similar to TF-IDF, BM25 is constrained by its dependence on lexical overlap, lacking the ability to capture semantic relationships between different expressions of the same concept. As a result, its performance diminishes when queries rely on paraphrases or contextual meaning.

2.2 Semantic Retrieval

The lexical limitations of TF-IDF and BM25 motivated the transition toward **semantic retrieval**, where the focus is on capturing the *meaning* behind words, sentences, and documents rather than just their surface form. Advances in natural language processing (NLP) have enabled models that embed text into high-dimensional spaces where semantically similar items are positioned close together.

Distributed Word Representations: Introduced by Mikolov et al. in 2013, Word2Vec was one of the first neural approaches to learning distributed word embeddings. It consists of two main architectures: Continuous Bag-of-Words (CBOW), which predicts a target word from its surrounding context, and Skip-Gram, which predicts the surrounding words given a target word. By training on large corpora, Word2Vec learns dense, low-dimensional vectors in which semantically related words such as king and queen are positioned close together, enabling analogical reasoning (e.g., king – man + woman ≈ queen). Compared to sparse representations, Word2Vec captures semantic similarity more effectively, is computationally efficient, and has been widely applied across natural language processing tasks. However, it produces static embeddings, assigning the same vector to a word regardless of its context, which limits its ability to handle polysemy (e.g., bank in finance vs. bank of a river). Despite these shortcomings, Word2Vec marked a major shift toward semantic retrieval and laid the groundwork for more advanced contextual embedding models.

Contextual Embeddings (BERT): The introduction of BERT revolutionised natural language processing by generating context-dependent embeddings that capture the meaning of words in relation to their surrounding text. Unlike Word2Vec, which produces static embeddings, BERT represents the same word differently depending on context, for example, the word cell is encoded distinctly in cell biology versus cell phone. BERT is pre-trained on two core tasks: Masked Language Modelling (MLM), which involves predicting masked words from context, and Next Sentence Prediction (NSP), which helps the model understand sentence-level relationships. This pretraining enables BERT to excel at downstream tasks such as question answering, named entity recognition, and semantic similarity, and in retrieval, it allows systems to interpret the semantic meaning of queries and documents rather than relying solely on lexical overlap. However, BERT also faces limitations: it is computationally expensive, requiring joint processing of queries and documents at inference time, and it primarily produces token-level embeddings, which are not directly optimised for sentence-level similarity tasks.

Sentence-BERT for Efficient Semantic Retrieval (SBERT): To address the inefficiencies of BERT in retrieval tasks, Reimers and Gurevych (2019) proposed Sentence-BERT (SBERT), which fine-tunes BERT using Siamese and triplet network architectures to generate fixed-size sentence embeddings. Unlike vanilla BERT, SBERT enables embeddings for documents to be pre-computed and efficiently compared using cosine similarity, making it highly scalable for large corpora. Its strengths include producing high-quality sentence-level embeddings that are well-suited for clustering, ranking, and semantic similarity, while significantly outperforming vanilla BERT on semantic textual similarity benchmarks. SBERT is particularly effective in document retrieval scenarios where queries are expressed in natural language and must be matched against semantically relevant documents, even in the absence of direct keyword overlap.

Multilingual Embeddings: In globally distributed information systems, multilingual retrieval is critical. Multilingual embeddings such as mSBERT and LaBSE (Language-agnostic BERT Sentence Embeddings) extend SBERT’s architecture to align semantic spaces across languages. This allows, for example, a query in English to retrieve documents in French or Hindi if they share semantic meaning. Such cross-lingual retrieval is vital for international research databases, government repositories, and enterprise systems operating across diverse linguistic environments. While multilingual models improve inclusivity and access, challenges remain in handling low-resource languages, cultural nuances, and domain-specific terminology.

The foundations of document retrieval reflect a clear progression from keyword-based to semantic approaches. Early methods, such as TF-IDF and BM25, established the groundwork for relevance ranking but were limited by their dependence on lexical overlap. The introduction of neural embedding models like Word2Vec, followed by contextual transformers such as BERT and its optimised variant SBERT, expanded retrieval beyond surface-level keywords by capturing semantic and contextual meaning. This transition from lexical to semantic retrieval represents a paradigm shift in information access, enabling systems that can interpret user intent rather than relying solely on exact term matching. However, it also introduces new challenges, including computational complexity, domain adaptation, and concerns about data privacy in semantic operations. A concise comparison of these methods, their strengths, limitations, and typical use cases is presented in Table 1, which highlights the trade-offs that continue to shape the design of modern retrieval systems.

Table 1: Comparison of Document Retrieval Methods

Method	Type	Strengths	Limitations	Example Use Cases
TF-IDF	Keyword-based	Simple, efficient	No semantics	Web search baseline
BM25	Keyword-based	Weighted scoring	Limited context	Digital libraries
Word2Vec	Semantic	Captures similarity	Context window limited	Embedding-based IR
BERT	Semantic	Deep context	Heavy compute	QA, legal retrieval
SBERT	Semantic	Sentence-level embeddings	Domain fine-tuning needed	Semantic search engines

3. Security and Privacy in Retrieval

As digital repositories continue to expand across sectors such as healthcare, finance, law, and government, the demand for secure document retrieval has intensified. Sensitive data stored in cloud environments or shared across multiple organisations must be accessible for legitimate purposes while remaining shielded from unauthorised access or inference attacks. Traditional keyword-based retrieval systems, even when augmented by modern semantic models like SBERT, often require access to plaintext queries and documents for indexing and similarity computation. This creates significant vulnerabilities in privacy-sensitive settings. To address these concerns, researchers have explored a variety of security mechanisms. This section reviews three broad categories: traditional encryption schemes, advanced searchable and homomorphic encryption techniques, and federated learning paradigms.

Traditional Encryption: Advanced Encryption Standard (AES): The Advanced Encryption Standard (AES) is a symmetric-key block cipher and has been the cornerstone of modern data protection for decades. It encrypts fixed-size blocks of data (typically 128 bits) using keys of length 128, 192, or 256 bits. Its main advantages are efficiency, widespread hardware acceleration, and resistance to known cryptanalytic attacks. AES is extensively used to secure documents at rest (e.g., files on servers) and in transit (e.g., HTTPS, VPNs). However, AES presents a critical limitation for information retrieval: once encrypted, data must generally be decrypted before being searched, indexed, or analysed. This requirement poses risks when decryption occurs in untrusted environments, such as third-party clouds. For example, suppose a medical institution stores patient records encrypted with AES on a cloud server. In that case, retrieval queries must either (1) download and decrypt entire datasets locally, which is computationally inefficient, or (2) allow decryption on the cloud, which compromises confidentiality. Thus, while AES is highly effective for storage and transmission security, it offers no native support for retrieval operations.

Rivest–Shamir–Adleman(RSA): The RSA cryptosystem is a widely used public-key algorithm based on the hardness of integer factorisation. It is often employed for secure key exchange, digital signatures, and encrypting small data blocks such as session keys. RSA is particularly important for establishing secure channels in client-server retrieval systems. Yet, similar to AES, RSA is not designed for large-scale encrypted search. It can protect communication between users and servers, but once documents are encrypted under RSA, they must be decrypted to enable operations such as similarity scoring or clustering. The computational cost of RSA encryption also makes it impractical for bulk document protection. Both AES and RSA provide strong guarantees for data confidentiality at rest and in transit, but fall short in enabling search or computation over encrypted data. This limitation paved the way for searchable encryption and homomorphic encryption schemes that aim to bridge the gap between security and functionality.

Searchable Encryption and Homomorphic Encryption: Searchable Encryption (SE): Searchable Encryption enables users to perform keyword searches directly on encrypted data, ensuring confidentiality while maintaining retrieval functionality. Broadly, it falls into two categories. Searchable Symmetric Encryption (SSE) uses a symmetric key to encrypt documents and construct an encrypted index, with queries transformed into encrypted tokens that allow the server to search without revealing plaintext content. While early SSE schemes supported only exact keyword matching, later advancements incorporated ranked retrieval and dynamic updates. In contrast, Public-Key Encryption with Keyword Search (PEKS) is an asymmetric approach where third parties can encrypt data for a recipient, who later searches using their private key. PEKS is particularly useful in scenarios such as secure email search, but remains less efficient for large-scale repositories. Despite its practicality, SE faces limitations: most schemes are still keyword-centric and cannot easily integrate semantic embeddings like SBERT vectors, access-pattern leakage and query repetition pose security risks, and extensions for ranked or fuzzy search often introduce significant computational overhead. Nonetheless, SE continues to provide a valuable balance between security and functionality, especially in enterprise and cloud-based keyword retrieval systems.

Homomorphic Encryption (HE): Homomorphic Encryption is a powerful cryptographic technique that enables computations to be performed directly on encrypted data, producing results that, once decrypted, match the outcome of operations performed on the original plaintext. Depending on the level of functionality, HE can be classified into three main types: Partially Homomorphic Encryption (PHE), which supports a single operation such as addition (Paillier) or multiplication (RSA); Somewhat Homomorphic Encryption (SHE), which allows limited additions and multiplications; and Fully Homomorphic Encryption (FHE), which supports arbitrary computations but remains computationally expensive. In the context of document retrieval, HE has been applied to tasks such as computing similarity scores (e.g., cosine similarity or dot products) over encrypted SBERT embeddings, enabling encrypted ranking, clustering, and classification, and supporting secure multiparty collaboration without exposing raw data. Despite its promise, HE faces several challenges: ciphertexts are significantly larger than plaintexts, homomorphic operations incur heavy computational overhead, noise accumulation during repeated operations requires costly bootstrapping, and securely performing top-k selection for retrieval remains complex.

Nevertheless, advances such as the CKKS scheme for approximate arithmetic have improved the practicality of HE, particularly for machine learning and retrieval applications where relative ranking is sufficient. As a result, HE is emerging as a key enabler of privacy-preserving semantic search, especially when integrated with embedding models like SBERT.

Federated Learning (FL): While Searchable Encryption (SE) and Homomorphic Encryption (HE) focus on enabling secure computations over encrypted data, another key challenge arises when training retrieval models on sensitive corpora, where traditional centralised training that aggregates raw data on a single server is infeasible due to regulatory and privacy constraints. Federated Learning (FL) addresses this challenge by enabling decentralised training across multiple clients, such as hospitals, banks, or law firms, where models are trained locally on private data. Instead of sharing raw documents, clients transmit model updates (e.g., gradients or weights) to a central aggregator, which combines them to produce a global model. This model is then redistributed to clients for further local training in an iterative cycle. For document retrieval, FL offers significant benefits: sensitive documents remain local to each institution, multiple organisations can collaboratively train SBERT-based encoders for semantic retrieval without exposing raw text, and models can be adapted to specific domains such as biomedical or legal corpora while preserving privacy. However, FL also faces notable challenges, including communication overhead due to large model updates, difficulties with data heterogeneity across clients, risks of privacy leakage through gradient inversion attacks, and vulnerabilities to adversarial behaviours such as model poisoning. To mitigate these risks, FL is often combined with secure aggregation protocols to prevent the server from viewing individual updates, differential privacy techniques to obscure contributions of individual records, and homomorphic encryption to ensure confidentiality during transmission.

Modern cryptographic techniques offer varying degrees of security and functionality for information retrieval. Traditional methods such as AES and RSA provide strong guarantees for secure storage and transmission, with AES enabling fast symmetric encryption and RSA supporting secure key exchange. However, neither offers native search capabilities over encrypted data. Searchable Encryption (SE) addresses this limitation by allowing efficient keyword search over ciphertext, making it practical for enterprise search. However, it remains restricted in semantic capability and susceptible to leakage of access patterns. Homomorphic Encryption (HE) extends the scope by enabling semantic computations, such as similarity scoring over encrypted SBERT embeddings, thereby ensuring strong privacy; yet its high computational overhead limits large-scale deployment. Federated Learning (FL) introduces a complementary paradigm by facilitating distributed training without requiring raw data sharing, enabling institutions to train retrieval models collaboratively. Nonetheless, FL necessitates safeguards such as secure aggregation, differential privacy, or homomorphic encryption to defend against gradient leakage and adversarial attacks. A structured comparison of these techniques, including their advantages, drawbacks, and common applications, is summarised in Table 2.

4. Optimisation in Retrieval

Efficient document retrieval requires not only robust semantic representations and privacy-preserving mechanisms but also effective optimisation strategies. As datasets grow in scale and retrieval models become more computationally intensive, there is a pressing need for algorithms that can fine-tune parameters, accelerate query processing, and balance competing objectives such as accuracy, latency, and security. Classical optimisation methods often fall short when dealing with the high-dimensional, nonlinear, and dynamic search spaces characteristic of information retrieval. To address these challenges, researchers have increasingly turned to metaheuristic optimisation techniques, inspired by natural processes and collective intelligence. Among the most prominent are Particle Swarm Optimisation (PSO), Genetic Algorithms (GA), Ant Colony Optimisation (ACO), and the Butterfly Optimisation Algorithm (BOA). These methods not only enhance retrieval performance but can also be adapted to privacy-preserving contexts, where traditional optimisation is constrained by encrypted or distributed data.

Table 2: Cryptographic Techniques for Secure Retrieval

Technique	Security Model	Advantages	Drawbacks	Applications
AES	Symmetric	Fast, secure	No search support	Data storage
RSA	Asymmetric	Secure key exchange	Slow for large data	Key management
Searchable Encryption	Keyword search over ciphertext	Efficient search	Limited semantics, leakage risk	Enterprise search
Homomorphic Encryption	Computation over ciphertext	Strong privacy	High computational overhead	Secure SBERT similarity
Federated Learning	Decentralised training	No raw data sharing, domain adaptability	Gradient leakage, adversarial risks	Collaborative model training

Particle Swarm Optimisation (PSO): PSO is a population-based stochastic optimisation technique inspired by the collective behaviour of bird flocks and fish schools. In this method, a group of particles explores the search space, with each particle representing a candidate solution. Particles iteratively update their positions by considering both their own best-known positions (personal best) and the best position identified by the entire swarm (global best). In the context of document retrieval, PSO has been applied to tasks such as optimising feature weights in ranking functions (e.g., tuning BM25 parameters), improving clustering of semantic embeddings (e.g., SBERT vectors), and performing hyperparameter optimisation for neural retrieval models. Its advantages include simplicity of implementation, few required hyperparameters, and strong effectiveness in continuous, high-dimensional optimisation spaces. However, PSO also faces limitations, including the risk of premature convergence to suboptimal solutions and performance sensitivity to parameter tuning.

Genetic Algorithms (GA): GA is an evolutionary optimisation technique inspired by the principles of natural selection, where candidate solutions are represented as chromosomes and iteratively improved through processes of selection, crossover, and mutation. Over successive generations, the population evolves toward solutions with higher fitness. In information retrieval, GA has been applied to query expansion by evolving sets of candidate keywords to maximise retrieval relevance, feature selection to identify optimal subsets of ranking or classification features, and optimising similarity thresholds in privacy-preserving retrieval environments. Its strengths lie in flexibility, the ability to explore large and discrete search spaces, and effectiveness at escaping local optima. However, GA can be computationally expensive when operating on large populations, and its performance depends heavily on the careful design of fitness functions.

Ant Colony Optimisation (ACO): ACO is a metaheuristic inspired by the foraging behaviour of ants, which use pheromone trails to communicate and collectively discover shortest paths. In computational adaptation, artificial ants construct solutions probabilistically, guided by both pheromone intensity and heuristic information, with stronger trails reinforcing more promising solutions over time. Within the domain of information retrieval, ACO has been applied to document clustering by simulating paths between documents where pheromone accumulation guides cluster formation, to query routing in distributed or federated retrieval systems, and to the optimisation of ranking functions under privacy-preserving constraints. Its advantages include being naturally suited for combinatorial optimisation tasks and its adaptability to dynamic environments where document collections evolve continuously. However, ACO is also limited by slow convergence rates and high sensitivity to parameters such as pheromone evaporation, which require careful tuning for effective performance.

Butterfly Optimisation Algorithm (BOA): The Butterfly Optimisation Algorithm is a relatively recent metaheuristic inspired by the foraging behaviour of butterflies, particularly their use of fragrance to navigate their environment. In BOA, each candidate solution is represented as a butterfly whose fragrance intensity is proportional to its fitness. Movement occurs either toward the current best solution (global search) or randomly toward other butterflies (local search), with the balance between these modes controlled by a switching parameter. In the context of information retrieval, BOA has been applied to tasks such as hyperparameter tuning for deep retrieval models like BERT and SBERT, secure optimisation of encrypted similarity computations, and balancing trade-offs between retrieval accuracy and computational overhead. BOA offers advantages in effectively balancing global exploration and local exploitation, while also showing stronger convergence properties compared to older metaheuristics such as Genetic Algorithms. However, its novelty means that empirical validation in retrieval applications is still limited, and in practice it may require hybridisation with other algorithms to enhance robustness..

Hybrid Metaheuristics in Secure Environments: While algorithms such as PSO, GA, ACO, and BOA each offer unique strengths, no single metaheuristic is universally optimal across all retrieval scenarios. This has led to growing interest in hybrid metaheuristics that combine complementary features of multiple algorithms, particularly in privacy-preserving retrieval systems. For instance, GA–PSO hybrids leverage GA's crossover and mutation to maintain population diversity while using PSO for faster convergence; ACO–PSO hybrids integrate the exploratory power of ant colonies with PSO's efficiency to improve document clustering; and BOA–GA hybrids combine BOA's balance of exploration and exploitation with GA's evolutionary adaptability. In secure retrieval environments, these hybrid approaches play a vital role by enabling optimisation under constraints such as encrypted data or federated settings where raw documents remain inaccessible. They can optimise similarity computations over homomorphically encrypted embeddings where computational resources are limited, fine-tune aggregation strategies in federated learning to balance accuracy with communication efficiency, and cluster encrypted embeddings without decryption by relying on metaheuristic-based distance approximations. By uniting the strengths of different paradigms, hybrid metaheuristics provide robustness and adaptability, offering efficient and privacy-aware solutions for modern retrieval challenges.

Table 3: Metaheuristic Optimisation Techniques in Retrieval

Algo.	Inspiration	Strengths	Limitations	IR Applications
PSO	Swarm intelligence	Fast convergence	Premature convergence	Feature weighting
GA	Evolutionary biology	Escapes local minima	Expensive computation	Query expansion
ACO	Ant foraging	Good for combinatorial tasks	Slow convergence	Document clustering
BOA	Butterfly foraging	Balanced search	Still new, less tested	Hyperparameter tuning

Optimisation plays a pivotal role in advancing document retrieval systems, especially when combined with modern semantic models and security frameworks. As shown in Table 3: Metaheuristic Optimisation Techniques in Retrieval, each metaheuristic algorithm brings distinct strengths and trade-offs. PSO provides simplicity and efficiency in high-dimensional optimisation, GA offers robust evolutionary search for feature selection and query expansion, ACO models collaborative behaviour for clustering and routing, and BOA contributes strong convergence through its bio-inspired mechanism. In practice, hybrid metaheuristics that combine these methods offer the most promising avenue, particularly for privacy-preserving retrieval environments where optimisation must navigate both computational and security constraints. The integration of metaheuristics with semantic models like SBERT, encryption schemes such as Homomorphic Encryption, and distributed learning frameworks like Federated Learning will be crucial in designing the next generation of retrieval systems. By harnessing nature-inspired optimisation, researchers can address the dual challenges of scalability and privacy, paving the way for secure, efficient, and intelligent information retrieval.

5. Research Gaps and Open Challenges

The development of privacy-preserving semantic document retrieval has made substantial progress in recent years, yet several fundamental challenges remain unresolved. These gaps highlight the inherent tension between accuracy and privacy, the difficulty of enabling clustering in encrypted spaces, and the complexities of combining federated learning with encrypted retrieval frameworks. Addressing these issues is crucial for the practical deployment of secure, large-scale retrieval systems.

Trade-off Between Privacy and Accuracy: A central challenge in privacy-preserving retrieval lies in achieving an effective balance between data confidentiality and retrieval performance. Encryption schemes such as Homomorphic Encryption (HE) and Searchable Encryption (SE) provide strong safeguards for sensitive information; however, they also introduce distortions or restrictions in similarity computation. For instance, in semantic retrieval with SBERT embeddings, cosine similarity is commonly employed to rank documents, yet once embeddings are encrypted, only approximate similarity computations are feasible. This often results in small but significant deviations in ranking outcomes, leading to reductions in precision and recall compared to plaintext retrieval. Similarly, in Federated Learning (FL), privacy is preserved through mechanisms such as differential privacy, where noise is injected into gradients or model updates. While these techniques are effective for protecting data, they can further diminish retrieval accuracy. In addition to accuracy, computational efficiency forms another dimension of this trade-off. Privacy-preserving approaches, particularly those based on HE, tend to be resource-intensive, creating latency in retrieval operations. For large-scale or time-sensitive applications such as legal document repositories or real-time healthcare decision support systems, even modest increases in computational cost can pose practical obstacles. Designing algorithms that deliver strong privacy guarantees without undermining retrieval effectiveness remains an open research problem. Promising directions include the development of approximate HE schemes (e.g., CKKS) that enable efficient vector computations, privacy-preserving distillation techniques that transfer knowledge from large semantic models into smaller ones while retaining utility, and adaptive privacy–accuracy mechanisms capable of dynamically tuning privacy levels based on the requirements of specific application contexts.

Lack of Clustering in Encrypted Space: Clustering plays a critical role in document retrieval, supporting tasks such as topic modelling, duplicate detection, and semantic organisation of large collections; however, performing clustering directly in encrypted space remains a significant challenge. Distance metrics such as Euclidean distance and cosine similarity are straightforward in plaintext, yet their secure computation over encrypted vectors is highly resource-intensive. Although Homomorphic Encryption (HE) enables certain arithmetic operations, iterative clustering algorithms like k-means require repeated distance calculations, centroid updates, and convergence checks, all of which compound computational overhead in encrypted settings. Searchable Encryption (SE) schemes are similarly constrained, as most are optimised for exact keyword matching rather than similarity-based grouping, and therefore lack mechanisms for forming semantic clusters or topic-based groupings over encrypted data. From a practical standpoint, this limitation hampers scalability: in large encrypted databases, clustering could dramatically reduce retrieval time by filtering candidate sets before fine-grained ranking, but without it, privacy-preserving systems must operate over entire corpora, resulting in inefficiency. Developing privacy-preserving clustering algorithms capable of functioning directly on encrypted embeddings remains an open challenge, with potential solutions including the design of approximate distance metrics that are computationally lighter under encryption, the application of secure multiparty computation (MPC) to distribute clustering tasks among semi-trusted parties, and the exploration of federated clustering approaches where local clusters are partially built and later aggregated securely.

Federated + Encrypted Retrieval Challenges: Federated Learning (FL) and encryption-based retrieval each provide strong privacy guarantees, yet their integration introduces several unresolved challenges. FL preserves locality by ensuring raw data never leaves the client. Homomorphic Encryption (HE) or Searchable Encryption (SE) secures model updates and queries; however, combining these approaches often amplifies existing bottlenecks. A primary issue is communication overhead: large neural architectures such as BERT or SBERT generate massive gradient updates, and encrypting these with HE further inflates their size, creating severe bandwidth and latency constraints when training across many clients. Moreover, federated gradients remain vulnerable to leakage through inversion attacks. While encryption can mitigate risks, the joint use of HE and FL requires lightweight protocols that maintain usability while defending against adversarial behaviours, including poisoning attacks where malicious clients manipulate encrypted updates. Another complication is semantic alignment across heterogeneous domains, for example, medical and legal corpora, where embeddings trained locally may diverge semantically, undermining interoperability in cross-client retrieval. Finally, performing top-k retrieval under FL and HE is computationally burdensome, as ranking encrypted vectors relies on either approximate similarity computations or expensive homomorphic comparisons that do not scale efficiently. Addressing these challenges requires the development of hybrid frameworks that balance communication efficiency, adversarial robustness, and semantic consistency. Promising research avenues include compression and quantisation of encrypted updates to mitigate communication costs, secure aggregation protocols that combine federated updates without exposing intermediate information, and cross-client embedding alignment mechanisms to ensure semantic interoperability across domains.

The research community has made notable strides in privacy-preserving semantic document retrieval, but several gaps remain unresolved (Table 4). First, the privacy–accuracy trade-off limits real-world deployment, as encrypted computations often degrade retrieval quality and introduce high latency. Second, the absence of practical clustering mechanisms in encrypted space prevents efficient organisation and scaling of large document repositories. Third, the combination of federated learning and encryption introduces new challenges in communication, adversarial robustness, and semantic interoperability. Addressing these open challenges requires interdisciplinary solutions, combining advances in cryptography, distributed systems, and natural language processing. By bridging these gaps, future research can enable retrieval systems that are simultaneously secure, scalable, and semantically effective, meeting the needs of sensitive domains such as healthcare, finance, and government intelligence.

Table 4: Summary of Research Gaps

Gap	Description	Potential Directions
Privacy–Accuracy Trade-off	Encryption reduces retrieval accuracy	Approximate HE, adaptive trade-offs
Clustering in Encrypted Space	Lack of secure similarity-based grouping	MPC, federated clustering
FL + Encrypted Retrieval	Overhead, gradient leakage	Secure aggregation, model compression

6. Conclusion and Future Directions

The pursuit of privacy-preserving semantic document retrieval lies at the intersection of semantic text representation, cryptographic security, and distributed learning. This survey has traced the progression from keyword-based models (TF-IDF, BM25) to advanced semantic embeddings (Word2Vec, BERT, SBERT), alongside security-preserving approaches such as AES, RSA, Searchable Encryption, and Homomorphic Encryption (HE). It has also highlighted the contributions of Federated Learning (FL) in enabling collaborative model training on sensitive datasets and metaheuristic optimisation (PSO, GA, ACO, BOA) in improving retrieval efficiency and accuracy. Several conclusions emerge. Semantic embeddings significantly outperform keyword methods, but their deployment in sensitive domains demands stronger privacy guarantees. Traditional cryptographic techniques protect data at rest or in transit but fall short for computation, a gap partially filled by SE and HE. HE and FL serve as complementary strategies: the former allows secure computation over encrypted embeddings, while the latter prevents raw data sharing across clients. Metaheuristics provide a practical means to balance speed, accuracy, and privacy. Despite these advances, challenges persist, including the privacy–accuracy trade-off, clustering in encrypted spaces, and efficient integration of FL with encrypted retrieval. Looking forward, hybrid frameworks that integrate HE, SE, FL, and differential privacy represent a promising direction. Advances in approximate HE schemes, cross-client semantic alignment, and lightweight secure clustering protocols are crucial for scalability. Adaptive privacy–utility mechanisms and explainable secure retrieval will further align research with real-world demands. By addressing these gaps, future systems can achieve both semantic richness and robust privacy, critical for domains such as healthcare, finance, and government intelligence.

Reference

[1]. Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys (CSUR), 51(4), 1–35.

[2]. Alaparthi, S., & Mishra, M. (2020). Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127.

[3]. Alhijawi, B., & Awajan, A. (2024). Genetic algorithms: Theory, genetic operators, solutions, and applications. Evolutionary Intelligence, 17(3), 1245–1256.

[4]. Assouline, L., & Minaud, B. (2023). Weighted oblivious RAM, with applications to searchable symmetric encryption. In Annual International Conference on the Theory and Applications of Cryptographic Techniques (pp. 426–455). Springer.

[5]. Brochier, R., Guille, A., & Velcin, J. (2019). Global vectors for node representations. In The World Wide Web Conference (pp. 2587–2593).

[6]. Carston, R. (2021). Polysemy: Pragmatics and sense conventions. Mind & Language, 36(1), 108–133.

[7]. Carvalho, T., Moniz, N., Faria, P., & Antunes, L. (2023). Towards a data privacy–predictive performance trade-off. Expert Systems with Applications, 223, 119785.

[8]. Choi, J. I., & Butler, K. R. B. (2019). Secure multiparty computation and trusted hardware: Examining adoption challenges and opportunities. Security and Communication Networks, 2019(1), 1368905.

[9]. Cui, J., Wu, L., Huang, X., Xu, D., Liu, C., & Xiao, W. (2024). Multi-strategy adaptable ant colony optimisation algorithm and its application in robot path planning. Knowledge-Based Systems, 288, 111459.

[10]. Fatima, S., Rehman, T., Fatima, M., Khan, S., & Ali, M. A. (2022). Comparative analysis of AES and RSA algorithms for data security in cloud computing. Engineering Proceedings, 20(1), 14.

[11]. Fu, J., Wang, N., Cui, B., & Bhargava, B. K. (2021). A practical framework for secure document retrieval in encrypted cloud file systems. IEEE Transactions on Parallel and Distributed Systems, 33(5), 1246–1261.

[12]. Garcia, M., Vieira, T. K., Scarton, C., Idiart, M., & Villavicencio, A. (2021). Probing for

idiomaticity in vector space models. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 3551–3564).

[13]. Ghorbel, A., Ghorbel, M., & Jmaiel, M. (2017). Privacy in cloud computing environments: A survey and research challenges. The Journal of Supercomputing, 73(6), 2763–2800.

[14]. Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

[15]. Hoblos, J. (2020). Experimenting with latent semantic analysis and latent Dirichlet allocation on automated essay grading. In Proceedings of the 2020 7th International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1–7). IEEE.

[16]. Hu, Z., Dai, H., Yang, G., Yi, X., & Sheng, W. (2022). Semantic-based multikeyword ranked search schemes over encrypted cloud data. Security and Communication Networks, 2022(1), 4478618.

[17]. Jain, S., Jain, S. K., & Vasal, S. (2024). An effective TF-IDF model to improve the text classification performance. In Proceedings of the 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 1–4). IEEE.

[18]. Juneja, S., Goyallal, S., Agarwal, S., Agrawal, S., Kumar, R., Dewang, R., & Mewada, A. (2022). Spam review detection using the Okapi relevance method for negative reviews. In Data, Engineering and Applications: Select Proceedings of IDEA 2021 (pp. 493–504). Springer.

[19]. Kannan, S. K., & Diwekar, U. (2024). An enhanced particle swarm optimisation (PSO) algorithm employing quasi-random numbers. Algorithms, 17(5), 195.

[20]. Mewada, A., & Dewang, R. K. (2023). SA-ASBA: A hybrid model for aspect-based sentiment analysis using synthetic attention in a pre-trained language BERT model with extreme gradient boosting. The Journal of Supercomputing, 79(5), 5516–5551. https://doi.org/10.1007/s11227-022-05159-9

[21]. Muqadas, A., Khan, H. U., Ramzan, M., Naz, A., Alsahfi, T., & Daud, A. (2025). Deep learning and sentence embeddings for the detection of clickbait news from online content. Scientific Reports, 15(1), 13251.

[22]. Nagy, B., Hegedűs, I., Sándor, N., Egedi, B., Mehmood, H., Saravanan, K., Lőki, G., & Kiss, A. (2023). Privacy-preserving federated learning and its application to natural language processing. Knowledge-Based Systems, 268, 110475.

[23]. Natarajan, N., & Venugopal, M. (2025). Hybrid butterfly optimisation and back propagation neural network for enhanced smart city data classification. Environmental Science and Pollution Research, 1–20.

[24]. Patel, V., Hiran, D., & Dangarwala, K. (2023). Recent trends of information retrieval systems: Review based on IR models and applications. In International Conference on Recent Trends in Machine Learning, IoT, Smart Cities & Applications (pp. 619–629). Springer.

[25]. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Association for Computational Linguistics.

[26]. Prakasha, S., Raju, G. T., & Singh, M. K. (2016). Cluster optimisation in information retrieval using self-exploration-based PSO. International Journal of Intelligent Engineering Informatics, 4(1), 91–115.

[27]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.

[28]. Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P. S., & He, L. (2024). Deep clustering: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems, 36(4), 5858–5878.

[29]. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389.

[30]. Sanjalawe, Y., Al-E'mari, S., Abualhaj, M., Makhadmeh, S. N., Alsharaiah, M. A., & Hijazi, D. H. (2025). Recent advances in the secretary bird optimisation algorithm, its variants and applications. Evolutionary Intelligence, 18(3), 65.

[31]. Shanmugam, V., Ling, H.-C., Gopal, L., Eswaran, S., & Chiong, W. R. (2024). Network-aware virtual machine placement using enriched butterfly optimisation algorithm in cloud computing paradigm. Cluster Computing, 27(6), 8557–8575.

[32]. Wang, Y., Wang, J., & Chen, X. (2016). Secure searchable encryption: A survey. Journal of Communications and Information Networks, 1(4), 52–65.

[33]. Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. Communication Methods and Measures, 15(2), 81–102.

[34]. Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., & Zhang, W. (2023). A survey on federated learning: Challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2), 513–535.

[35]. Zhou, Y., Shi, Y., Zhou, H., Wang, J., Fu, L., & Yang, Y. (2023). Toward scalable wireless federated learning: Challenges and solutions. IEEE Internet of Things Magazine, 6(4), 10–16.