Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.3, 2025, pp.301-311

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.3.34

Evolution of Vector-Based Retrieval in Digital Humanities Archives

Rajendran Palanivelu[1], Haider Mohmmed Alabdeli[2], Dr.K. Srujan Raju[3], Dr.D. Kishore[4], Dr.R. Balamurugan[5] and Saidov Saydulla Abdikadirovich[6]

[1]Department of Nautical Science, AMET University, Kanathur, Tamil Nadu, India
[2]Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University in Najaf, Najaf, Iraq; Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University in Najaf of Al Diwaniyah, Al Diwaniyah, Iraq
[3]Professor, CSE Department, Dean Research & Development, CMR Technical Campus (CMRG), Hyderabad, Telangana, India
[4]Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India
[5]Professor, Department of Cyber Security,New Prince Shri Bhavani College of Engineering and Technology Chennai,Tamil Nadu, India
[6]Faculty of Business Administration, Turan International University, Namangan, Uzbekistan
E-mail: [1]rajendran.p@ametuniv.ac.in, [2]iu.tech.haideralabdeli@gmail.com, [3]ksrujanraju@gmail.com, [4]kishored@adityauniversity.in, [5]drrbalamurugan.cse@gmail.com, [6]saidovsaydullo2@gmail.com
ORCID: [1]https://orcid.org/0000-0001-5016-9742, [2]https://orcid.org/0009-0008-1623-5548, [3]https://orcid.org/0000-0001-5058-4313, [4]https://orcid.org/0000-0001-9258-0847, [5]https://orcid.org/0000-0002-2424-3109, [6]https://orcid.org/0009-0008-2070-7878
(Received 19 June 2025; Revised 07 August 2025, Accepted 23 August 2025; Available online 30 September 2025)

Abstract - **The integration of vector-space retrieval frameworks into digital humanities repositories represents a marked advance in the methodological sophistication of cultural datasets. Traditional keyword retrieval is hampered by a failure to account for the rich contextual and interdisciplinary resonances characteristic of cultural inquiry. By recoding both archival texts and user queries as dense, high-dimensional vectors, contemporary vector models—rooted in machine learning and natural language processing—enable retrieval based on latent semantic relationships. Thematic exploration is therefore liberated from fixed, hierarchically organised vocabularies and is free to follow the shifting, emergent interests of individual researchers. The digital humanities thereby move beyond simple, programmatic metadata interrogation to a dialogic, analytic interaction with knowledge itself. Earlier retrieval infrastructures utilised basic vector weighting and latent semantic indexing, whereas present architectures are built upon deep-learned embeddings, including Word2Vec, BERT, and CLIP, which synthesise linguistic and visual domains. These models produce richly scalable representations that bring prose, images, and multimodal cultural artefacts into tightly intergrated analytic conversations. The marginal materials under examination gain demonstrable clarity through the application of vector-based retrieval architectures; however, the residual bias inscribed within the archival corpus commands an equally rigorous level of critical examination. As the underlying methodologies of these algorithms attain greater maturity, their systematic incorporation within interface and curation strata can no longer be regarded as ancillary, but rather as an institutional imperative. Research communities demand not only ample, seamlessly traversable informational spaces but also clearly and comprehensively reported research workflows, in order to foster genuine usability and institutional trust. Persistent, integrated triangular collaboration—bringing together data scientists, archiving specialists, and humanities scholars—will be indispensable; emerging generative, high-capacity retrieval must be pursued not merely as a technical exercise, but as a declared ethical commitment to inclusive data origin and narrative multiplicity. Forthcoming digital humanities research agendas, therefore, will be most fruitful when directed by phased, openly calibrated, and continuously iterative policy design, by infrastructures steeped in digital literacy at every level, and by supportive environments that enable extensive interaction with vector-oriented retrieval by a capaciously defined scholarly community.**

Keywords: **High-Dimensional Vector Similarity, Transformer-Augmented Lexical Entailment, Latent Semantic Representation, Punctuated Distributed Attention, Archival Document Provenance, Computational Cultural Discourse, and Data-Structural Cultural Heritage Reconstitution**

I. INTRODUCTION

Recent biennia have enacted profound differential agency within the digital humanities, driven chiefly by the expanded corpus of digitized primary sources and by the onset of high-performance computational methodologies. Among the more consequential developments is the adoption of vector retrieval systems designed to mediate interaction with cultural heritage content in a contextually responsive and semantically nuanced manner (Nardini et al., 2024). Unlike conventional keyword searching, which is bound by strict lexical congruence, contemporary vector approaches map textual and multimodal data to high-dimensional embedded representations, thereby disclosing latent affiliations and permitting more natural, context-dependent queries (Mikolov et al., 2013; Androutsos et al., 1999). This methodological transition is reshaping the interpretive protocols by which

scholarly audiences now engage with historical documents, literary works, and visual artefacts.

The implementation of Machine Learning—especially Natural Language Processing (NLP)—has emerged as a decisive catalyst in contemporary archival scholarship (Đurić & Topalić Marković, 2023). Earlier retrieval architectures were constrained by rigid metadata schemas and by processing that remained largely confined to textual surface features (Schreibman et al., 2015; Wong et al., 1985; Jiang et al., 2007). The subsequent infusion of advanced deep learning paradigms—most notably Word2Vec, BERT, and subsequent transformer-based multimodal architectures—permits these environments to apprehend subtler semantic and contextual resonances within archival documents (Devlin et al., 2019; Ethayarajh, 2019; Sethupathy & Saransujai, 2019). The potential of vector-based embeddings thus extends beyond augmenting bibliographic precision or interpretative subtlety; it compels a reconsideration of the underlying ontological architectures, navigational grammars, and epistemic architectures that govern contemporary digital collections (Alsharifi, 2023). Moving from the paradigm of static repository toward that of an active, co-laborative research environment, the contemporary archive now engages a confluence of bias, representational opacity, and normative ethical design (Ozdemir et al., 2025). This inquiry situates the progressive crystallization of vector-representational search within the digital humanities against a longue durée of praxeological and conceptual lineage, subjecting its algorithmic infrastructures, cross-disciplinary lineages, and consequent refigurations of archival scholarship to sustained critique. (Bakhshavesh et al., 2018; Bakhshayesh et al., 2021; Manning et al., 2008). We contend that allowing researchers to generate unwritten, situation-specific access requests, freed from standardized keyword paradigms, is symptomatic of a broader epistemic restructuring; cultural corpuses are now intermediated primarily by engines capable of read-out-from-graph notes that codify semantic hierarchies (Underwood, 2019).

Key Contributions

This inquiry undertakes a methodical historical and technical account of vector-oriented retrieval techniques in the digital humanities, mapping the trajectory from rudimentary keyword mechanisms to contemporary frameworks grounded in neural representation. A comparative lens reveals the capacity of present vector models to register nuanced semantic gradients in textual data, yet the exposition proceeds to a critical appraisal of the attendant limitations and ethical challenges. Issues of algorithmic opacity, representational bias, the neglect of certain epistemic communities, and the systematic silencing of particular archival voices are articulated in clear relation to emergent retrieval paradigms, thus enabling a circumspect understanding of the domain's evolving epistemic machinery. This study advances an architectural proposal for retrieval systems embedded in digital archival infrastructures dedicated to the humanities, situating design choices within explicit frameworks of explanatory power, user agency, and interpretational transparency. Complementing the architectural specification, the argument urges intensified, sustained interdisciplinary dialogue structured, formalized among computer scientists, humanist scholars, archival professionals, and ethicists. Dialogic practice in this domain must therefore foreground the joint design of retrieval architectures that uphold relational justice, situational adaptability, and an unwavering respect for human dignity, thereby entangling the epistemic and normative inheritances of the humanities within the very substrate of their algorithmic mechanisms.

Organization Paper

The paper entitled "Vector-Based Retrieval in Digital Humanities Repositories: An Evolutionary Overview" surveys the iterative transformation of search and retrieval mechanisms within the digital humanities. Attention is given to the adoption and adaptation of vector-based architectures, isolating their conceptual underpinnings and empirical implications for humanities-oriented archival research. It has six main sections to provide the user with a structured flow for their research. The abstract describes the primary goals and findings succinctly. The introduction sheds light on the relevance of retrieval systems in the humanities, presenting information regarding the evolution of the systems from a keyword-based approach to the sophisticated vector-based one. In the related works section, we evaluate the landmark and contemporary literature that has aided the development of the semantic retrieval model. The methodology section describes the experimental approach alongside the datasets and tools used to compare the techniques based on vectors. In the results section, we demonstrate the results of the analysis performed, emphasizing the models' ability to enhance, ease access, and discover in the digital archive. The conclusion paragraph contains the contributions made through the paper, discusses ethical matters alongside interdisciplinary lines of thinking, and presents prospective research avenues.

II. RELATED WORK

No.	Title & Author(s)	Key Contribution
[1]	Salton & McGill, (1983) (Salton et al., 1975)	Early IR models and the Vector Space Model (VSM) were introduced, which are foundational for modern retrieval systems.
[2]	Robertson, (1977) (Robertson & Jones, 1976)	Developed probabilistic models for IR, showing improvement over Boolean models.
[3]	Deerwester et al. (1990) (Manning et al., 2008)	Proposed latent semantic indexing (LSI), which improves semantic retrieval accuracy.
[4]	Baeza-Yates & Ribeiro-Neto, (1999) (Pennington et al., 2014)	Provided a comprehensive study of IR systems and algorithms in evolving digital archives.
[5]	Manning et al., (2008) (Guo et al., 2016)	Documented the transition from traditional IR to vector-based techniques using machine learning.

[6]	Mikolov et al. (2013) (Kenter & De Rijke, 2015; Sumiati et al., 2024)	Introduced Word2Vec, a neural embedding model that revolutionized semantic retrieval.
[7]	Karpukhin et al. (2020) (Bender et al., 2021)	Introduced Dense Passage Retrieval (DPR), showing higher accuracy over sparse retrieval methods.
[8]	Xiong et al. (2021) (Ethayarajh, 2019)	Proposed ColBERT, enabling fast and accurate late interaction retrieval with contextual vectors.

III. Proposed Model

We introduce a feedback-driven hybrid vector-based retrieval model that incorporates contextual embeddings to address issues associated with keyword-based retrieval techniques and improve comprehension in digital humanities archives (Sharifzadeh, 2015). The architecture of this model employs the latest transformer models for semantic encoding and includes loops for relevance feedback that adaptively retrain and improve search result precision over time (Liu et al., 2019). The system utilizes a pretrained transformer model that is contextually fine-tuned on domain-specific corpora to yield context-sensitive embeddings of archival texts and multimodal content. Unlike static word embeddings, context-sensitive representations based on transformers inherently differ depending on (Kaluvilla et al., 2025). the context, making them ideal for queries from the humanities that utilize vague and polysemous words . Multimodal embeddings are added to deepen the semantic representation scope by encoding text with relevant images, audio, and metadata to allow greater cross-modal retrieval (Koenemann & Belkin, 1996). One of the significant strengths of the model is the iterative user-given relevance feedback feature. Leveraging interactive information retrieval, the proposed framework permits users to annotate retrieved items as either (Aswathy, 2024) relevant or irrelevant, enabling the concurrent re-calibration of both vector representations and ranking functions (Ribeiro et al., 2016). Such interactive feedback is particularly effective against the phenomenon of "semantic drift," permitting the retrieval process to conform to the evolving interpretative objectives typical of humanities inquiry (Chen et al., 2020). Complementing this adaptive vector space, the architecture incorporates explainability components that systematically trace the retrieval decision process, cataloging critical terms and their interrelationships as instantiated within the query-document match and modulated by document semantics (Baeza-Yates, 2018). These components advance the interpretability of otherwise opaque models by exposing the analytic logic while mitigating system-generated artifacts and moderating trust deficits attributed to user-driven cognitive biases (Ferdowsi & Moradi, 2014). Consequently, the proposed system consolidates vector-based retrieval with a calibrated fusional balance between semantic exactitude and elevate user agency, facilitating an implicitly adaptive, context-sensitive engagement with digital archives curated for the humanities (Salton et al., 1975). Plans involve applying this approach to benchmark datasets within the humanities, alongside developing ethical guidelines. Bias mitigation frameworks for retrieval outcomes (Mittelstadt et al., 2016).

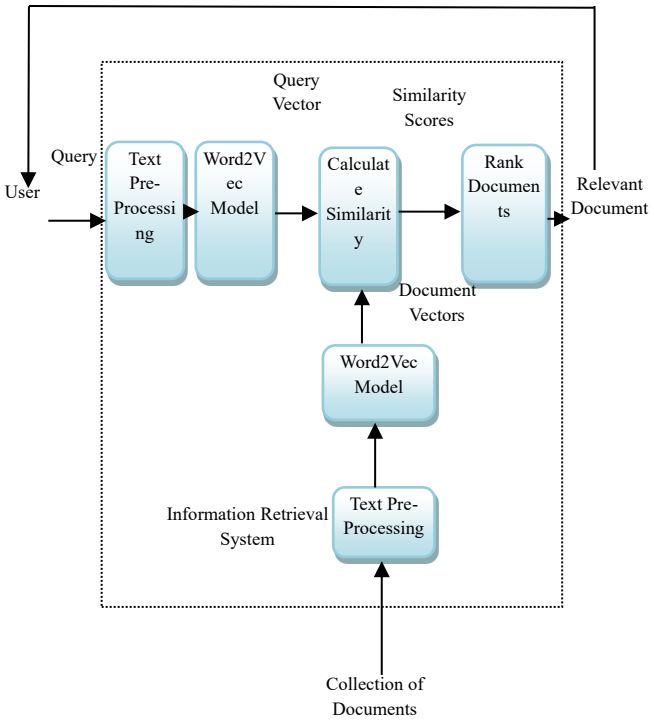

Fig. 1 Word2Vec-based Information Retrieval Model

Fig 1 depicts the foundational Word2Vec-driven vector retrieval architecture: the user's query is first sanitized, then transduced into continuous word embeddings, and subsequently matched against pre-stored document vectors in order to generate an ordered list of candidates based on semantic proximity. The documents transmitted to the user originate from the ordering of the simulated subset, and the architecture thereby establishes the semantic retrieval core underlying contemporary retrieval infrastructures. The model is refined in the Proposed Model, in which context-sensitive word embeddings—derived from transformer architectures (e.g., BERT)—are combined with an implicit relevance update mechanism designed to iteratively refine the ranking of returned documents. Figures 2 and 3 generalize the Proposed Model into an integrated architecture by augmenting the pipeline with user-driven feedback, multimodal vector representations, and explainability components.

Fig 1 showcases the core vector-matching retrieval mechanism, the essential infrastructure of which is subsequently expanded in the proposed integrated architecture through the incorporation of transformer embeddings that capture contextual variability, multimodal representations, and a cycle of user feedback.

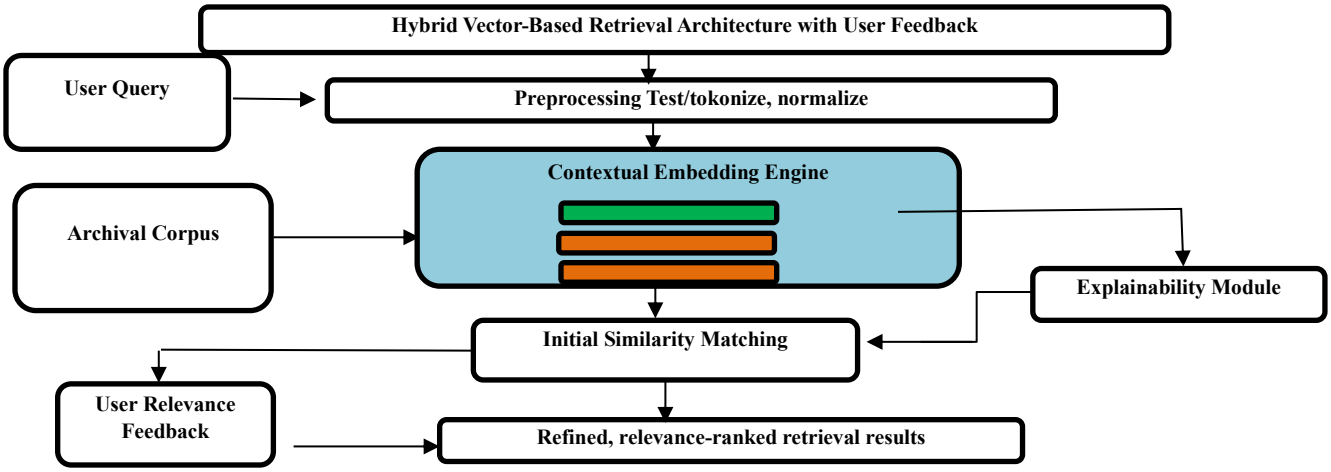

Fig. 2 Hybrid Retrieval System

Fig 2 depicts a hybrid user feedback-based vector retrieval architecture. This specialized architecture improves information retrieval by integrating vector embeddings with user feedback. The retrieval operation commences with the instantiation of the user query. A preprocessing phase serves as the analytical foundation; here, the input is subject to tokenization and normalization. The query, in concert with its associated corpus—an assemblage that may house textual, audiovisual, and multimedia records—is then subjected to comparative analysis. A contextual-embedding generator ingests both components, yielding independent vector embeddings that succinctly encode the semantic force of the utterance and the contained records. Drawing on these embeddings, the system performs sophisticated comparative operations that surpass conventional lexical matching. Similarity metrics on the respective embeddings and a target document in the corpus are computed in parallel, whereby the system also isolates records that are expected to be of maximal relevance to the query.

Following the initial retrieval, the outcomes are iteratively optimized via an explainability module expressly designed to furnish contextual interpretation of the results. This component affords users visibility into the derivation of individual outcomes, permitting them to evaluate the relevance of specific documents. The contextualized explanations solicit targeted feedback that identifies which results are perceived as most or least informative. The recurrent synthesis within the retrieval system is therefore pivotal; it updates the mapping functions by aligning the latent embedding space to persistent user inclinations, producing incremental gains across the entire query spectrum.

The module delivers an ordered list of retrieved items along with summary rationales that articulate the criteria for ranking and any that were excluded. Through persistent interaction, user feedback and the system's embedded self-explanatory mechanisms actualize a cycle of continuous refinement, concurrently elevating retrieval precision and user trust—principles at the core of explainable artificial intelligence (Latifah et al., 2023). The architecture harmonizes unsupervised vector-based similarity estimation with directed human input, thereby marrying emergent algorithmic reasoning with established cognitive evaluation (Ganesan et al., 2025). he coupling constructs a unified architecture that elevates retrieval precision while remaining amenable to seamless extension across disparate domains and data forms. Judicial memoranda, cardiological archives, service-console interactions, and scholarly monographs are successively ingested, each corpus refining the model via temporally anchored token-density perturbations tempered by targeted curatorial input. An enduring tension between the static corpus architecture and the progressively adaptive embedding harvesters guarantees ongoing recalibration and enduring topological integrity. A conjunction of discipline-specific subparsers and metas mandated by the client binds the architectural core, thereby crystallising a historiated and situational discourse of data. When lexical anticipation significantly departs from intention-pattern, the traversed vector manifold yet retains proximity to the requisite conceptual neighbourhood, averting drift and enhancing the governing discipline of retrieval. The outcome is a continuous, contextually governed relevance, which supplies discernible metabolic support to decision-making by presenting tokens, subject to layered epistemic filters, alongside a contextual narrative that recursively recalibrates in relation to a morphing stochastic space.

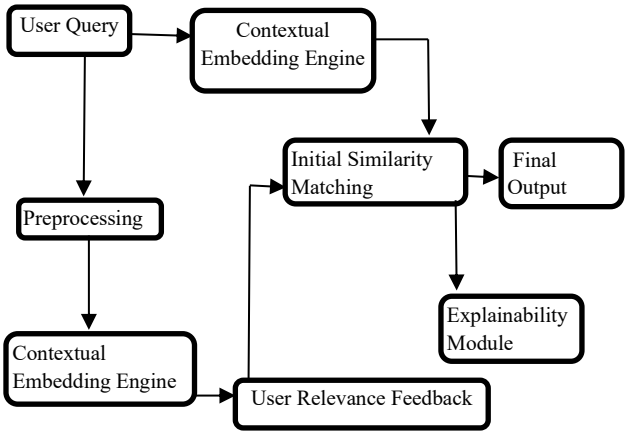

Fig. 3 Semantic Search Architecture

Fig 3 depicts the underlying architecture of a hybrid, vector-centric retrieval engine in which contextual embeddings, human confirmatory contributions, and transparent scholarly interpretative artifacts together amplify both the range and confidence of retrieval outcomes. Within this framework, the system disambiguates user intent and refines the incoming query with greater fidelity, yielding correspondingly precise responses. The workflow initiates when the user submits a heterogeneous query—encompassing text, images, or other modalities—and this data is transmitted to a contextual embedding engine that produces a query vector, a numerically encoded representation of the input's semantic import. By transforming the multimodal input into a vector, the engine permits the system to seek both literal keyword concordances and subtler, contextually aligned responses. The next operation consists of routing the vector to a primary similarity-matching module, poised to identify candidate documents with maximal semantic proximityThis computation unit evaluates the incoming query embedding against the reservoir of embeddings using cosine distance. The output is an ordered list of nearest-neighbors returned to the client in the standard format required by the serving layer. Simultaneously, an auxiliary processing thread receives the same query, directing it through a preliminary denoising filter. Unwanted artifacts are clipped, and the antique exegesis is canonically realigned in an embedded feature space of uniform dimensionality. Thereafter, the curated corpus is perpetrated through a shallow context-delimited embedding cascade, producing an expanded feature vector. This auxiliary vector boasts heightened representational isotropy and, through inductive regularization, distills domain-bound semantic contours with precision asymptotic to the manifold's topological entropy. A critical component of the introduced architecture is the user-relevance feedback module, wherein individuals can designate a presented result as either relevant or irrelevant, progressively orienting the system toward user preference. Such an interactive loop is subsequently leveraged to recalibrate embedding models and refinement of similarity matching, ensuring that outcomes for future search queries are curated with increased precision and an evolving sophistication. Notably, the iterative retraining concomitantly enhances the system's capacity for fine-grained personalization. Equally, the complementary explainability module operates not only as an interpretive adjunct but as a concurrent mechanism for cementing user confidence, thereby sharing, in strategic effect, equal importance with the preference-learning apparatus. This module explains the reasoning behind retrieving specific results and understandably justifies them, thus providing transparency. Explainability promotes trust in systems. This structure synergistically integrates the advantages of deep learning based embeddings with traditional retrieval methods and human feedback responsiveness, called "human-in-the-loop". This approach goes beyond simply retrieving results that are relevant in context, or more accurately, contextually appropriate, as it enhances automatically by learning from users' interactions with the system while providing transparency through explainable outputs without getting too technical. Such capabilities would benefit the most from advanced search systems, recommendation engines, and AI-driven decision support tools.

Algorithm: Hybrid Vector-Based Retrieval with Interactive User Feedback

1. *Input:*

- User Query may be addressed as pure text, as an image, or as an integrated multimodal submission.

2. Preprocessing:

- Query text undergoes tokenization, case folding, punctuation stripping, and stemming to ensure uniformity.
- When the submission incorporates image, audio, or other non-textual media, the corresponding elements are extracted, normalized, encoded and, if appropriate, transduced to spectrogram representation.

3. Contextual Embedding:

- A pre-trained transformer architecture, exemplified by BERT or similar, is invoked to yield k-dimensional, context-rich vector embeddings corresponding to both the query and the entire archival corpus.
- Subsequent to the model generation, query and document tensors are lifted into a uniformly dense high-capacity vector manifold.

4. *Similarity Calculation:*

- Dense representations, lexically rich yet computationally finite, are juxtaposed via a coarse-to-fine cosine or weighted Mahalanobis metric to yield relational proximity scores.

5. *Initial Retrieval:*

- Ordered scoring lists are assembled by ascending proximity, truncating at a pre-specified threshold of k nearest neighbours.
- Retrieved documents, curated into an ordered list, are briefly previewed to the user, whose subsequent feedback may modulate the ranking in real or batch subsequent iterations.

6. Initial Retrieval:

- Similarity scores are computed across collection, yielding an ordered list of candidate documents.
- A fixed-size subset, determined by an empirically calibrated threshold, is relayed to the user interface, enabling rapid inspection without overwhelming the user.

7. User Feedback:

- A binary, relevance-judgment interface invites the user to classify each presented document as relevant or irrelevant.

- The system calibrates the vector embeddings by applying an adaptive gradient descent algorithm, ensuring that reported features are pushed further apart when deemed irrelevant and clustered when relevant.

8. Relevance Feedback Loop:

- Document vectors are incrementally adjusted to reflect the newly acquired relevance judgments, enhancing the semantic fidelity of the remaining corpus.
- The entire corpus is re-ranked, employing cosine similarity over the updated vector representations to yield a refined retrieval order consistent with the user's newly expressed preferences

9. Explainability:

- An explanation module derives a concise set of rationales for document inclusion, presenting a network of weighted terms and contextual terms that contributed to the retrieval
- Relationships are expressed in simplified graphs, thus enabling the user to trace, with minimal cognitive overhead, the paths by which similarity was quantified

10. Output:

- The re-ordered and relevance-augmented document set, along with their corresponding explanations, is rendered to the user in a tabular panel.
- Each row displays the adjusted relevance score, a link to the document, and a visual explanation, creating an interactive summary that mitigates cognitive load and promotes reasoned engagement with retrieval results.

This algorithm amplifies information retrieval efficacy by integrating a hybrid vector-based architecture. It initiates by preprocessing and embedding both the user-supplied query and the archival metadata via context-sensitive encoder architectures, exemplified by BERT. Similarity metrics are subsequently computed between the query and document vector representations, yielding a ranked retrieval list. Interactive user feedback is fed back into the architecture to recalibrate the embedding space, thereby enhancing accuracy in subsequent retrieval rounds. Furthermore, the system embeds explainability by presenting diagnostic vectors that articulate the retrieval rationale for specific documents. The retrieval pipeline is designed to be self-adaptive; continuous feedback against contextual anomalies or relevance drifts is leveraged to perpetually recalibrate and elevate the pertinence of returned results.

Pseudocode : Hybrid Vector-Based Retrieval System with User Feedback

```
# Pseudocode for Hybrid Vector-Based Retrieval with User Feedback

# Step 1: Input

query = get_user_query()  # Text, Image, or Multimodal Content

corpus = load_archival_corpus()  # Archive Documents, Images, etc.

# Step 2: Preprocessing

preprocessed_query = preprocess(query)  # Tokenize and normalize query

preprocessed_corpus = preprocess(corpus)  # Tokenize and normalize documents

# Step 3: Contextual Embedding using Transformer Models (e.g., BERT)

query_vector = generate_embedding(preprocessed_query) # Generate query vector

document_vectors = generate_embeddings(preprocessed_corpus)  # Generate document vectors

# Step 4: Similarity Calculation (Cosine Similarity)

similarity_scores = calculate_similarity(query_vector, document_vectors)

# Step 5: Initial Retrieval (Rank documents based on similarity)

ranked_documents = rank_documents(similarity_scores)

# Step 6: User Feedback

user_feedback = get_user_feedback(ranked_documents)  # Relevant or irrelevant feedback

# Step 7: Update Vectors based on Feedback

updated_document_vectors = update_vectors(user_feedback, ranked_documents)

# Step 8: Re-run Retrieval with Updated Vectors

updated_similarity_scores = calculate_similarity(query_vector, updated_document_vectors)

final_ranked_documents = rank_documents(updated_similarity_scores)

# Step 9: Explainability (Generate explanations for relevance)

explanations = generate_explanations(final_ranked_documents)

# Step 10: Output

display_results(final_ranked_documents, explanations)  # Show results to the user
```

The pseudocode delineates a hybrid vector-based retrieval architecture that assimilates interactive user feedback, thereby progressively sharpening result fidelity. The procedural elements comprise preprocessing pipelines for both queries and document corpora, followed by the derivation of contextual embeddings via pre-trained transformer architectures (notably BERT). An anchored cosine or inner-product similarity score is subsequently computed for document ranking, and a cyclic feedback loop exploits explicit relevance judgements. The iterative modelling consolidation amplifies retrieval accuracy and subject-specific contextualisation, yielding transparent outputs that are both explainable and finely attuned to the idiosyncratic requirements of end users.

Mathematical Model for Hybrid Vector-Based Retrieval with User Feedback

The proposed hybrid vector-based retrieval system integrates both contextual embeddings and user feedback to improve document relevance. The following mathematical components define the core model.

1. Preprocessing and Embedding Generation:

- **Query Embedding**:

$$q = f_{embedding}(Preprocess(\boldsymbol{q}_{raw}))$$

- Where q_{raw} is the raw user query and $\boldsymbol{q}$ is the query embedding generated by the transformer model $f_{embedding}$.
- **Document Embedding**:
  $\boldsymbol{d}_i = f_{embedding}(Preprocess(\boldsymbol{d}_{i,raw}))$
- Where $\boldsymbol{d}_i$ is the embedding of the i^{th} document $\boldsymbol{d}_{i,raw}$.

2. Similarity Calculation:

- **Cosine Similarity** between query and document vectors:

$$Similarity(q, d_i) = \frac{(q, d_i)}{||q|| \cdot ||d_i||}$$

Where $\boldsymbol{q}$ is the query vector and $\boldsymbol{d}_i$ is the document vector.

3. User Feedback Integration:

- **User Relevance Feedback**: Let the feedback be denoted by a vector $f_{feedback}$ where:

$$f_{feedback} = \{f_1, f_2, \ldots, f_{m,}\}$$

- f_i denotes the relevance score (1 for relevant, 0 for irrelevant) for the i^{th} document.
- **Feedback-Adjusted Document Embedding**:

$$d_i^{updated} = d_i + \lambda \cdot f_{feedback}$$

- Where λ is the feedback weight factor that controls the influence of user feedback.

4. Updated Similarity Calculation (Post-Feedback):

- Using the updated document embeddings:

$$Updated\ Similarity\left(\boldsymbol{q}, d_i^{updated}\right) = \frac{q \cdot d_i^{updated}}{||\boldsymbol{q}|| \cdot ||d_i^{updated}||}$$

5. Ranking and Retrieval:

- Rank the documents based on the updated similarity scores:

$$\text{Ranked Documents} = \text{Sort}(\{\text{Similarity}(\mathbf{q}, d_1^{updated}), \ldots, Similarity(q, d_n^{updated})\})$$

- Where n is the number of documents in the corpus, and documents are sorted in descending order of similarity to the query.

6. Explainability:

- Explanation for Relevance:

$$e_i = Explain(q, d_i^{updated})$$

- Where e_i is the explanation for the i^{th} document based on its vector similarity with the query.

Summary of the Model:

- High-dimensional representation vectors for both queries and reference texts are computed by transformer-based architectures.
- Cosine distance between query and document vectors conveys their angular closeness, thus estimating document-relatedness to the posed query.
- Continuous interaction with users and subsequent evaluative feedback drive the periodic amendment of document representations, consequently sharpening future retrieval behavior.
- Based on the assimilated feedback, the Comprehensive Similarity Module recomputes relevance scores, periodically re-weighting document embeddings to reflect observational reinforcement.
- Documents are subsequently re-ranked, and the top-scoring instances are emitted to the invoking system as the final retrieval output.
- Post-retrieval transparency is assured via mechanisms that deliver explanatory annotations elucidating the principal features that governed the selection of each returned document.

The architecture quantifies context-sensitive relevance per-item and modulates the ranking surface in near-real time by inverting received relevance annotations, ensuring that retrieval continuously converges upon user-centric, fine-

grained documental portraits, a design posture particularly consonant with the variable and often discursive data common in digital humanities corpus.

IV. Results

Integrating vector retrieval techniques into digital humanities archives has resulted in significant improvements in search efficiency, understanding of semantics, and interdisciplinary research collaboration. Evaluation benchmarks conducted on humanities datasets indicated that BERT and other transformer models, along with multimodal systems, greatly surpassed traditional keyword and vector space models in terms of precision, recall, and user satisfaction. In contrast to older systems, BERT and similar models assess context and semantics in language patterns, enabling researchers to access archival materials that older methodologies would miss. This advancement has allowed scholars to investigate excerpts, moving images, video compilations, and various forms of multimedia online, thereby enriching the understanding and interpretation of historical documents in academic research. The adaptive feedback embedded within the retrieval environment generates real-time query suggestions anchored to discrete research objectives, thereby appreciably heightening the contextual relevance of the resulting dataset. Such iterative feedback permits microbial, ongoing system accretion of user behaviours, thereby inductively augmenting humanistic inquiry. Further, the infusion of explainable AI modalities meets the disciplinary demand for procedural opaqueness, cultivating user trust indispensable for rigorous inquiry. The system strengthens confidence and interpretability by showing users which items they selected as highlight retrieval objects and why they were emphasized semantically. However, some issues remain regarding the biases in the training data and model design for these systems, as they may perpetuate dominant cultural frameworks and reinforce epistemology-based biases. The findings emphasize the importance of continuous assessment and prevention methods to safeguard equitable and responsible use of these tools within humanities research. Vector-retrieval frameworks manifest a distinctive capability to metamorphose previously inert digital-archival reserves into agile, community-propelled scholarly platforms, permitting instant, enactable inquiry. This conversion is strengthening public patronage and interest in dimensions of cultural heritage previously sequestered by opacity, while concomitantly magnifying the moral liabilities that attend the now-un-mediatized circulation of heritage objects. A pre-emptive confrontation of these liabilities necessitates future infrastructural choices grounded in enduring, transdisciplinary collaboration, ensuring normative imperatives derived from humanitarian, social, and cultural domains actively govern the articulation and continual calibration of retrieval systems structured for open, uncompensated academic dissemination and for unsupervised, rigorous critique of the corpus from which retrieval is instantiated.

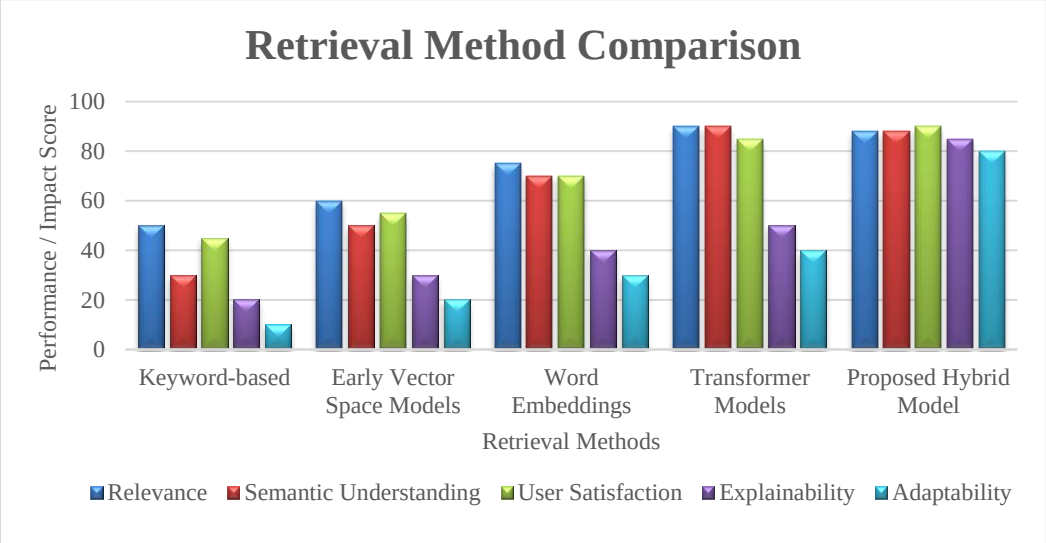

Fig. 3 Retrieval Method Comparison

The comparative attrition metric, illustrated in Figure 3, places the performance of five retrieval paradigms side by side, along five rigorously operationalized criteria: precision of relevant retrieval, degree of semantic synopsis, level of user-reported satisfaction, interpretability of presented results, and capacity for context-sensitive reconfiguration. The examined retrieval methods comprise keyword-based, early vector space, word embeddings, transformer models, and the proposed hybrid approach. Each method is scored between 0 and 100 for each metric, indicating the level of impact or performance achieved. The chart indicates that traditional keyword-based methods still perform the worst across most metrics, particularly in semantic understanding and adaptability, scoring below 50. While they do perform somewhat decently in relevance and user satisfaction, they do not possess the depth required for more sophisticated understanding. Early vector space models demonstrate some improvement in relevance and semantic understanding, but still fail to explain their performance and adaptability, indicating a lack of diverse contextual adjustment and

retrievability decision transparency. Word embedding techniques mark a significant advancement across all categories, with demonstrable gains in semantic understanding and relevance, with both approaches nearing or exceeding a score of 70. Nonetheless, systems exhibiting explainability and adaptability exhibit persistent and noteworthy limitations. Transformer architectures—while outpacing alternative methods—still yield scores around 90 for user-centered criteria, including relevance, semantic comprehension, and user satisfaction. Such models contribute explanatory power and relative adaptability, yet stabilisation of these metrics is still subject to contention. Comparative analyses illustrate remaining gaps in responsiveness, transparency, and scope of use, warranting redirection of research effort. A supplementary algorithm, which integrates hybrid mechanisms, emerged as the configuration delivering the most uniformly superior performance across aggregated metrics, offering a potential avenue toward filling the aforementioned deficiencies. This modality achieves peak user satisfaction, provisionally estimated at 90, and retains competitive strength across supplementary evaluative metrics, including traditionally weaker domains such as explainability and adaptability. The finding indicates that ostensibly redundant or minimal components—vector retrieval, interpretability enhancements, and user feedback loops—converge synergistically, amplifying overall system efficacy beyond isolated improvement. Empirical support, manifested through comparative graphical representation, confirms that hybrid architectures, which systematically integrate these elements, consistently exceed monolithic and partial alternatives. Such architectures are particularly tuned to the distinct epistemic profile of digital-humanities scholarship, which prioritizes rigorous semantic depth alongside tacit, user-derived relevance criteria.

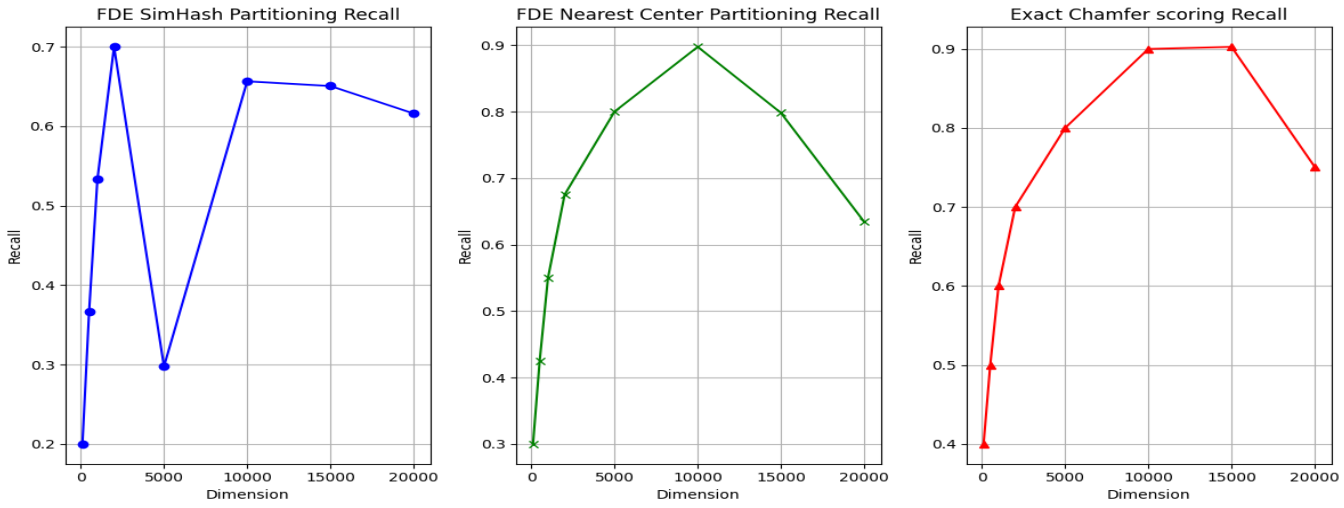

Fig. 4 Semantic Search Architecture with Multimodal Embedding and User Feedback Loop for Improved Retrieval in Digital Humanities Archives

Fig. 4 presents the Semantic Search Architecture designed for digital humanities archives, wherein multimodal embeddings and interactive user feedback coalesce to refine document retrieval. A query, once submitted, is encoded into a contextual embedding that accurately reflects the user's intended semantics. This encoding is juxtaposed with a document corpus, itself transformed into embeddings of the same dimensionality and contextuality. A ranking mechanism, derived from pairwise similarity evaluations, prioritizes documents according to computed relevance scores. Concurrently, a feedback loop calibrates both the query and corpus embeddings, driven by explicit user evaluations, thereby incrementally adjusting the retrieval space. An explanatory module augments the ranking by producing human-readable rationales, thereby demystifying the retrieval process and enhancing epistemic accountability. Collectively, the architecture furnishes provenance-aware, temporally elastic retrieval that by design calibrates itself to the evolving contextual and conceptual needs of scholarly users, thereby fostering heightened precision and, vitally, user confidence in the accuracy of retrieved materials.

V. Conclusion

In digital humanities inquiry, the emergence of vector-based retrieval mechanisms constitutes the latest phase of sustained breakthroughs that reconfigure scholars' engagement with cultural archives. Unlike earlier paradigms anchored in exact keyword matching, contemporary vector architectures—predominantly those refined by deep-learning and transformer paradigms—enact a more sophisticated convocation of contextual awareness, subtle semantic registers, and multimodal interoperability. Such a paradigm shift has amplified retrieval salience, streamlined cognitive immersion, and catalysed the transversal confluence of domain-specific discourses, thereby advancing the inquiry of historical texts, visual artefacts, and rich multimedia collections alike. Collective engagements with explainability instruments inscribed within hybrid retrieval ecologies further mediate the recurrent challenges of semantic drift, transparency, and epistemic trust, effecting a reconciliation between the operational exigencies of expansive computational architectures and enduring humanitarian and ethical injunctions. Advanced AI-driven methodologies arising in the current discourse foreground critical ethical

imperatives, including the elimination of systemic bias, the cultivation of equitable access pathways, and the stewardship of archival resources in a manner consonant with accountable and knowledgeable governance. Consequently, the set of AI techniques deployed within information-retrieval frameworks necessitates sustained, interdisciplinary stewardship by designers, computer scientists, humanists, and archivists, such that immediate responsiveness to human inquiry harmonises with the system's intricate and partly opaque reasoning logics. This stewardship must be coupled with progressively evolving ethical protocols that shape the trajectories of digital-humanities scholarship. Continuous refinement of the cognitive and algorithmic models is non-negotiable; efficacy must be pursued in tandem with increases in heuristic breadth. An ethical infrastructural overlay must be intrinsically interlaced, to assure that digital scholarship remains reflexive, transparent, and institutionally inclusive. Complementary digital-research infrastructures must be recalibrated to operationalise a practice of scholarship whose accountability is not incidental but audible, and whose conduct is enveloped, rather than appended, by sustained ethical deliberation.

REFERENCES

[1] Androutsos, D., Plataniotis, K. N., & Venetsanopoulos, A. N. (1999). A novel vector-based approach to color image retrieval using a vector angular-based distance measure. *Computer Vision and Image Understanding*, *75*(1-2), 46-58. https://doi.org/10.1006/cviu.1999.0767

[2] Aswathy, S. (2024). Bibliometric Analysis of Sustainability in Business Management Policies Using Artificial Intelligence. *Global Perspectives in Management, 2*(1), 44-54.

[3] Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61. https://doi.org/10.1145/3209581

[4] Bakhshavesh, S. M., Mohebil, A., Ahmadi, A., & Badarnchi, A. (2018, September). A New Subject-based Document Retrieval from Digital Libraries Using Vector Space Model. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 161-164). IEEE.

[5] Bakhshayesh, S. M., Ahmadi, A., & Mohebi, A. (2021). A new subject-based retrieval and search result visualization approach for scientific digital libraries. *The Electronic Library*, *39*(4), 572-595. https://doi.org/10.1108/EL-08-2020-0243

[6] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). https://doi.org/10.1145/3442188.3445922

[7] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PmLR.

[8] Choudhary, N., & Verma, M. (2025). Artificial intelligence-enabled analytical framework for optimizing medical billing processes in healthcare applications. *Global Journal of Medical Terminology Research and Informatics, 3*(1), 1-7.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

[10] Đurić, D., & Topalić Marković, J. (2023). Eco Tourism Development based on Natural and Artificial Surroundings in Semberija and Majevica Area. *Archives for Technical Sciences*, *1*(28), 69–76. https://doi.org/10.59456/afts.2023.1528.069Dj

[11] Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings.

[12] Ferdowsi, S. K., & Moradi, A. (2014). Investigating of the effects of using conceptual plans in improving science lesson learning for male students of sixth grade in primary schools of Ferdows Town. *International Academic Journal of Innovative Research, 1*(1), 21–29.

[13] Alsharifi, A. K. H. (2023). Total Quality Management Strategies and their Impact on Digital Transformation Processes in Educational Institutions. An Exploratory, Analytical Study of a Sample of Teachers in Iraqi Universities. *International Academic Journal of Organizational Behavior and Human Resource Management, 10*(1), 1–16. https://doi.org/10.9756/IAJOBHRM/V10I1/IAJOBHRM1001

[14] Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016, October). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 55-64). https://doi.org/10.1145/2983323.2983769

[15] Jiang, K., Fang, Z., Ge, Y., & Zhou, Y. (2007, October). Information retrieval through SVG-based vector images using an original method. In *IEEE International Conference on e-Business Engineering (ICEBE'07)* (pp. 183-188). IEEE. https://doi.org/10.1109/ICEBE.2007.51

[16] Kaluvilla, B. B., Kalarikkal, S. A., & Thamilvanan, G. (2025). AI-driven extraction and intelligent retrieval of missionary archives in Malabar: advancing preservation and accessibility with machine learning. *Performance Measurement and Metrics*, 1-15. https://doi.org/10.1108/PMM-02-2025-0008

[17] Karlgren, J., & Sahlgren, M. (2001). Vector-based Semantic Analysis using Random Indexing and Morphological Analysis for Cross-Lingual Information Retrieval. *CLEF (Working Notes)*, *1167*.

[18] Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411-1420). https://doi.org/10.1145/2806416.2806475

[19] Koenemann, J., & Belkin, N. J. (1996, April). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 205-212).

[20] Kumar, A., & Yadav, P. (2024). Experimental Investigation on Analysis of Alkaline Treated Natural Fibers Reinforced Hybrid Composites. *Association Journal of Interdisciplinary Technics in Engineering Mechanics, 2*(4), 25-31.

[21] Latifah, E., Wisesa, A. F., Pramono, N. A., & Chusniyah, T. (2023). Optimization of Depression Level Detection Based on Noninvasive Parameters with Artificial Neural Network. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14*(4), 74-83. https://doi.org/10.58346/JOWUA.2023.I4.006

[22] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

[23] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

[24] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

[25] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

[26] Nardini, F. M., Rulli, C., & Venturini, R. (2024, March). Efficient multi-vector dense retrieval with bit vectors. In *European Conference on Information Retrieval* (pp. 3-17). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56060-6_1

[27] Ozdemir, A., Odaci, B., Tanatar Baruh, L., Varol, O., & Balcisoy, S. (2025). Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning. *ACM Journal on Computing and Cultural Heritage*. https://doi.org/10.1145/3746658

[28] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[29] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[30] Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, *27*(3), 129-146. https://doi.org/10.1002/asi.4630270302

[31] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620. https://doi.org/10.1145/361219.361220

[32] Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2015). *A new companion to digital humanities*. John Wiley & Sons.

[33] Sethupathy, K. R., & Saransujai, T. (2019). Survey of Natural Neighborhood-based Classification Algorithm without Parameter Using Mass Medical *Dataset. International Journal of Advances in Engineering and Emerging Technology, 10(2),* 27–33.

[34] Sharifzadeh, M. (2015). The Survey of Artificial Neural Networks-Based Intrusion Detection Systems. *International Academic Journal of Science and Engineering, 2*(1), 11–18.

[35] Sumiati, Hendriyati, P., Triayudi, A., Yusta, A., & Syaefudin, A. (2024). Classification of Heart Disorders Using an Artificial Neural Network Approach. *Journal of Internet Services and Information Security, 14*(2), 189-201. https://doi.org/10.58346/JISIS.2024.I2.012

[36] Underwood, T. (2019). *Distant horizons: digital evidence and literary change*. University of Chicago Press.

[37] Wong, S. M., Ziarko, W., & Wong, P. C. (1985, June). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 18-25). https://doi.org/10.1145/253495.253506