OPEN ACCESS
Iḥyāʾalʿulūm
ISSN (Online): 2663-6263
ISSN (Print): 2663-6255
www.joqs-uok.com

Iḥyāʾalʿulūm, Vol.:25, Issue,1: Jan-June 2025
PP: 01-20

A Conceptual Framework for a Semantic Hadith Retrieval System Using Modern NLP Techniques

Sheikh Adnan Ahmed Usmani
Ph.D Scholar - Computer Science, Federal Urdu University of Arts Science and Technology,
Karachi, Pakistan.
Email: a2usmani@gmail.com

Wahyu Abdul Jafar
Universitas Islam Negeri Jurai Siwo Lampung, Metro, Indonesia.
Email: wahyujafar@metrouniv.ac.id

Abstract:

In the contemporary digital era, Islamic scholarship is undergoing a technological transformation, particularly in the realm of textual access and interpretation. One of the most critical corpora in Islamic studies is the Hadith literature, comprising the sayings, actions, and approvals of the Prophet Muhammad (Peace Be Upon Him). Traditional keyword-based search systems have proven inadequate for retrieving Hadith content effectively due to linguistic complexity, semantic variation, and contextual depth. This paper proposes a conceptual framework for a Semantic Hadith Retrieval System using modern Natural Language Processing (NLP) techniques. By integrating ontology-based thematic categorization, deep language models, and semantic similarity algorithms, the proposed framework aims to overcome the limitations of surface-level keyword matching. The paper begins by exploring the unique challenges posed by Hadith texts, followed by a comprehensive literature review of prior digital Hadith retrieval efforts and semantic search models. It then presents the theoretical foundations of semantic processing, outlines the ontological modeling required for religious themes, and delves into the technical components of NLP suitable for Arabic and Islamic contexts. Furthermore, a practical architectural framework is proposed, detailing the technology stack, implementation flow, and sample annotation practices. Ethical concerns, challenges of authenticity, and evaluation metrics are also discussed. This research does not involve the development of a functioning system but lays down a blueprint for future implementation. The framework is designed to guide both developers and Islamic scholars toward building more intelligent, spiritually aware, and user-friendly Hadith retrieval platforms. The integration of semantic web technologies with Islamic knowledge promises to not only modernize access to religious texts but also preserve their interpretive richness and contextual integrity.

Keywords: *Hadith Retrieval, Semantic Search, Natural Language Processing, Ontology-Based Categorization, Arabic NLP, Semantic Web, Islamic Scholarship*

1. Introduction

The Hadith literature represents one of the foundational sources of Islamic jurisprudence, theology, and spiritual guidance, second only to the Qur'an. Compiled over centuries, Hadiths encompass the sayings, actions, and tacit approvals of the Prophet Muhammad (peace be upon him), and have been meticulously preserved through chains of narration (isnad) and textual analysis (matn)[1]. The significance of Hadiths extends across various Islamic disciplines, including law (fiqh), ethics (akhlaq), theology ('aqidah), and social norms[2]. Their interpretive role in understanding the Qur'an and formulating Islamic rulings underscores the necessity of accurate, accessible, and contextually aware retrieval mechanisms[3].

In traditional settings, scholars relied on memorization, canonical texts, and personal networks to access Hadiths[4]. However, with the rise of digital libraries and databases, Hadith literature has been digitized and made searchable[5]. Despite this progress, most existing tools operate on rudimentary keyword-matching algorithms that fail to grasp the semantic nuances and linguistic diversity inherent in classical Arabic texts[6]. Consequently, users — whether scholars, students, or laypersons — often struggle to locate relevant narrations unless they possess exact phrasing or specialized knowledge. Given the explosion of interest in artificial intelligence (AI) and natural language processing (NLP) across disciplines, Islamic studies stands at a crossroads: it can either adopt these innovations to facilitate more profound engagement with its texts or risk remaining inaccessible to future generations. This paper asserts that semantic search technologies offer a critical pathway forward, especially when designed with sensitivity to Islamic epistemology and linguistic traditions[7]. A conceptual framework is necessary to bridge the gap between classical religious scholarship and modern computational linguistics.

2. Literature Review

The field of semantic search and retrieval, especially as applied to religious texts like Hadith literature, has garnered significant attention in recent years due to advancements in Natural Language Processing (NLP) and Information Retrieval (IR) technologies. This literature review surveys relevant academic and technical works across several domains: traditional Hadith retrieval systems, semantic information retrieval, ontological modeling in religious studies, and recent progress in neural-based language models applied to Islamic texts[8].

Traditionally, Hadith retrieval systems have focused on building searchable

databases using either keyword-based or Boolean search techniques[9]. Early platforms like Shamela, Maktaba Shamela[10] [11], and Hadith collections from Al-Islam.org primarily relied on exact text matching or syntactic filters[12]. These platforms, while user-friendly, suffered from poor semantic understanding and limited support for morphological diversity inherent in Arabic. For example, different derivations of the same root word could result in a mismatch if not explicitly encoded into the search system. These limitations made them suitable for scholars familiar with the specific terminology but less accessible to general users seeking meaning-based exploration.

The emergence of semantic search systems, largely driven by the need for contextual understanding in legal, biomedical[13], and academic[14] domains, has influenced work in religious text analysis as well. Semantic retrieval systems aim to go beyond the surface form of the words, focusing instead on latent meaning, synonyms, paraphrases, and contextual usage. Vector Space Models[15] (VSM), Latent Semantic Indexing (LSI), and more recently, neural embeddings such as Word2Vec[16], GloVe[17], and FastText, have laid the groundwork for capturing semantic relationships in textual corpora. These models convert words or documents into high-dimensional vectors that reflect contextual similarity, which has improved recall and precision in search applications. When applied to religious texts, these models enable retrieval of conceptually relevant content even in the absence of lexical overlap[18].

A key advancement in religious IR is the use of ontologies to model thematic, legal, and ethical dimensions of scriptural texts. In the context of Hadith, ontologies help classify narrations based on themes such as prayer, charity, ethics, or social justice. Studies by researchers like Omar Al-Ubaydli and others have attempted to formalize these thematic categorizations using RDF, OWL, and SKOS standards. Ontology-driven systems allow for semantic annotation and structured reasoning, which can enrich retrieval by linking narrations with similar meanings or shared jurisprudential implications. However, the construction of domain-specific ontologies for Hadith remains underdeveloped and often lacks consensus from Islamic scholarship[19]. More recently, the use of transformer-based language models such as BERT, RoBERTa, and their Arabic adaptations (AraBERT[20], CAMeLBERT[21], and AraELECTRA[22]) have revolutionized NLP for Arabic and Islamic texts. These models are capable of capturing fine-grained contextual dependencies and can be fine-tuned for semantic similarity, classification, and question-answering tasks. Research has shown that fine-tuned versions of BERT on religious texts[23] outperform traditional models in retrieving semantically similar hadiths or Quranic verses.

Efforts to represent Islamic knowledge as linked data on the Linked Open Data (LOD) cloud have largely focused on the Quran, with some recent initiatives exploring hadith modeling. These initiatives primarily annotate hadith components, such as Fairouz et al.[24], who model hadith commentaries, and Jaafar and Che Pa[25], who focus on concepts within Arabic hadith texts. Khatib et al.[26] proposed developing a WordNet linguistic resource for hadith, while Harrag et al.[27] applied association rule mining to construct an ontology for Islamic jurisprudence (Fiqh) from hadith texts. Similarly, Al-Arfaj and Al-Salman[28] introduced a framework for creating ontologies from Arabic texts. Other studies, such as those by Azmi et al.[29], review computational and natural language processing techniques applied to hadith literature, while works like[30] [31] focus on extracting hadith for indexing and classification. Saeed et al.[32] analyzed narrator chains as a social network of hadith narrators. Azmi and Bin Badia[33] developed E-Narrator, a tool that uses a natural language lexer for shallow parsing to generate narration chain trees and visually represent them through a graph presenter. Building on HadithRDF from E-Narrator[34], Baraka and Dallooul[35] created an ontology-based system that automatically suggests judgments on hadith isnad based on rules established by hadith scholars. Altammami et al.[36] employed text segmentation to annotate hadith texts, developed an Arabic-English hadith corpus, and suggested that a Quran ontology could serve as a foundation for modeling hadith knowledge. Despite these advancements, no publicly available linked data sources or vocabularies for hadith exist. Prominent hadith repositories, such as sunnah.com, Shamela Library, Encyclopedia of Harf: the Nine Books, and Aldourar Alsunnia, enable browsing and searching but lack standardized numbering schemes for hadith. The authenticity of hadith varies across sources, complicating their formal mapping and classification. Additionally, hadith texts differ in length, and commentators or Tafseer scholars often reference only portions of a hadith, making it challenging to trace the original source.

The presence of shared narration segments across multiple hadith further complicates accurate identification. A formalized knowledge representation and linking framework using linked data standards offers a promising approach to addressing these challenges.

Despite these advancements, gaps remain. Most existing works are either language-specific, focusing on Arabic-only or English-only corpora, or lack comprehensive integration between ontologies, semantic embeddings, and user intent modeling. Additionally, ethical considerations and the interpretative plurality of Islamic scholarship are seldom accounted for in the design of these systems. Therefore, this research proposes a hybrid framework that integrates semantic embeddings, ontology-driven categorization, and

multilingual capabilities to enable a deeper, context-aware retrieval experience for Hadith literature.

3. Methodology

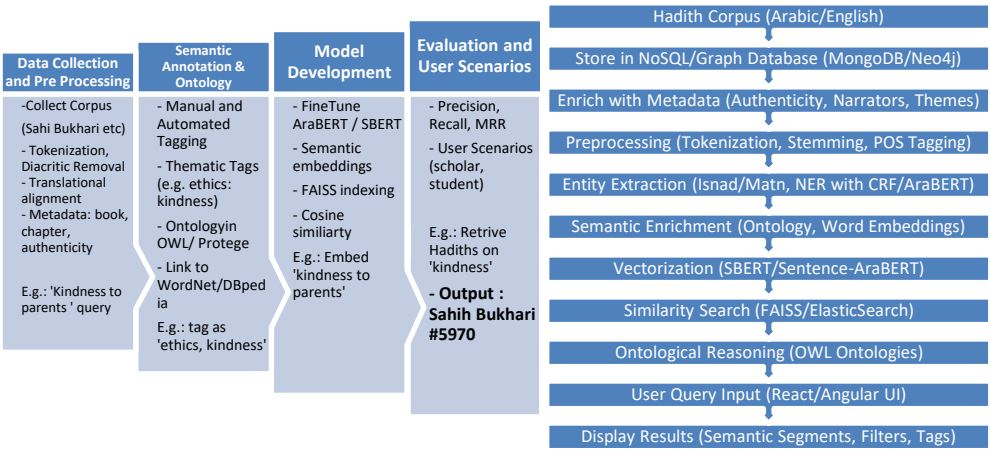

Figure 1: Proposed Framework for Semantic Hadith Retrieval System

3.1 Data Collection and Preprocessing

The first step in building a semantic Hadith retrieval framework is the compilation of a well-curated, authentic corpus of Hadith literature. This includes classical collections such as Sahih Bukhari, Sahih Muslim, Sunan Abu Dawud, Jami` at-Tirmidhi, and others. Publicly available digital repositories[37] such as Al-Maktaba Al-Shamila[38] and various Hadith APIs were considered. However, since the reliability of digital sources varies, cross-verification with authenticated printed versions and scholarly-approved digital libraries was emphasized.

Once the corpus was acquired, the next challenge was preprocessing the Hadith data to prepare it for semantic enrichment. This involved several sub-steps: tokenization, sentence segmentation, removal of diacritics, and normalization of Arabic orthography. Due to the morphologically complex nature of Arabic[39], classical NLP tools like Farasa[40] and CAMeL Tools[41] were evaluated for morphological analysis and part-of-speech tagging. Additionally, translation alignment between Arabic and English versions of the Hadith was conducted using parallel corpora, ensuring semantic consistency between languages. Figure 2 shows the isnad (chain of narration) and matn (text of the Hadith) were also separated during this phase to allow for independent semantic treatment.

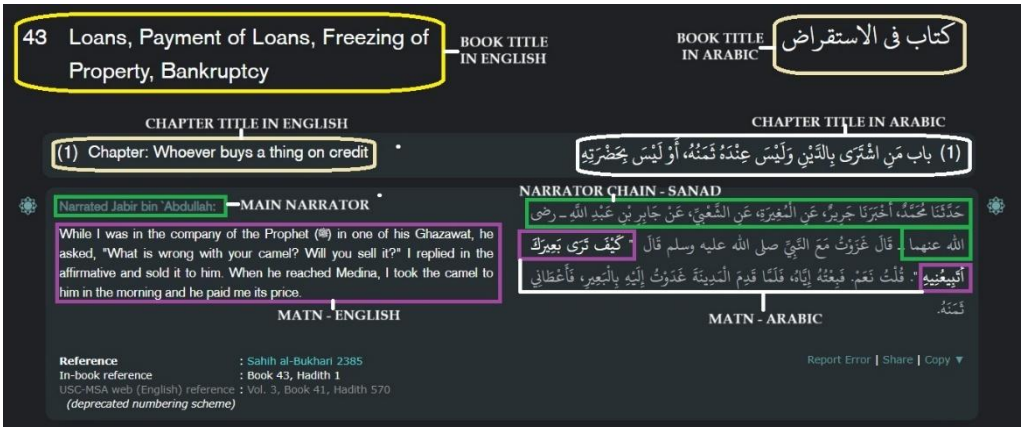

Figure 2: A sample snapshot of a hadith. This is taken from sunnah.com for the first hadith of the Sahih Bukhari collection. The snapshot shows the structure of the hadith and highlights the important components.

To ensure the integrity of the data, a metadata schema was introduced. Each Hadith entry was associated with metadata fields such as book source, chapter title, grading of authenticity, narrator chain, and thematic tags. This structured representation laid the groundwork for subsequent semantic annotation and retrieval layers. The preprocessing step culminated in a cleaned, structured, and multilingual Hadith corpus ready for semantic enrichment.

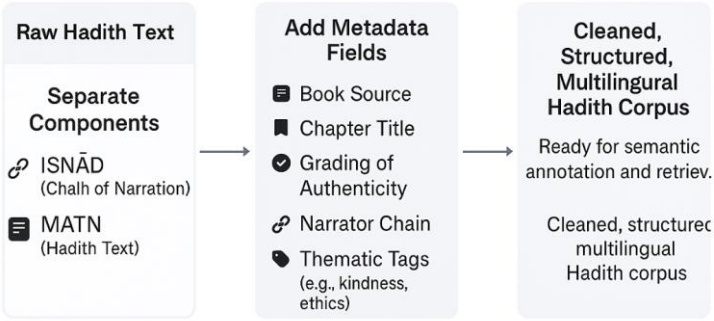

Figure 3: Preprocessing pipeline of the Hadith corpus showing the separation of ISNĀD and MATN, metadata enrichment, and generation of a cleaned, structured, multilingual corpus ready for semantic annotation and retrieval.

3.2 Semantic Annotation and Ontology Construction

Semantic annotation involves labeling the Hadith corpus with tags that reflect its underlying meaning rather than just its surface form. This was achieved

using a combination of manual tagging by Islamic scholars and automated annotation via NLP models trained on similar textual domains. Semantic annotation fields included thematic domains (e.g., jurisprudence, ethics, eschatology), emotional tone (e.g., mercy, fear, hope), and syntactic function (e.g., command, narrative, analogy).

A key part of this phase was the construction of a domain-specific ontology[42]. This ontology served as a structured vocabulary that described the concepts and their interrelations within Hadith literature. The ontology was developed using OWL (Web Ontology Language) and Protégé[43], based on pre-defined taxonomies derived from Islamic jurisprudence (fiqh), ethics (akhlaq), and theology (aqeedah). There are many Ontologies based editors available online amongst which a comparision is made by Altarish[44]Relationships such as "isPartOf," "isNarratedBy," "isGradedAs," and "hasTheme" were defined to enable sophisticated semantic querying. Moreover, multilingual labels in Arabic and English were added to each ontology node to support cross-lingual retrieval.

The ontology also included links to external semantic resources such as Arabic WordNet and DBpedia[45] where relevant, enhancing interoperability. This structured semantic backbone made it possible for the retrieval engine to interpret queries based on conceptual mappings rather than mere keyword overlaps, setting the stage for a contextually aware Hadith search system.

3.3 Model Development for Semantic Search

The next step was to develop a semantic retrieval model that could process user queries and match them with relevant Hadiths based on conceptual similarity. To achieve this, transformer-based language models such as AraBERT[46], mBERT[47], and SBERT (Sentence-BERT) [48] were explored. These models were fine-tuned using a task-specific dataset comprising user queries and relevant Hadiths annotated by domain experts.

Embedding techniques were used to convert both the Hadith text and user queries into dense vector representations. These embeddings captured semantic meaning and contextual relationships, enabling the use of cosine similarity and other distance metrics for retrieval. SBERT was particularly effective due to its capability to generate sentence-level embeddings optimized for semantic similarity tasks. A dual-encoder architecture was implemented, where one encoder processed the query and another processed the Hadith texts. For indexing and retrieval, FAISS[49] (Facebook AI Similarity Search) was used to allow fast nearest neighbor searches over the high-dimensional vector space. This made the retrieval engine scalable and

responsive, capable of handling thousands of Hadiths in real-time. The model was tested using a validation set of queries, with precision, recall, and mean reciprocal rank (MRR) being the key evaluation metrics.

3.4 Evaluation Planning and User Scenarios

To validate the effectiveness of the semantic retrieval framework, an evaluation plan was established. This included both intrinsic and extrinsic evaluations. Intrinsic evaluation would be focused on model performance using traditional IR metrics such as precision at k, recall, and F1-score. A sample test set of user queries and expert annotators labeled the correct Hadiths to serve as ground truth. Extrinsic evaluation, on the other hand, was user-centric. Several user scenarios were designed to reflect real-world use cases. For example, a scholar looking for legal precedents regarding "interest" in transactions, a student searching for hadiths on "kindness to parents," and a general user asking about "anger management" from an Islamic perspective. In each scenario, the system's ability to return relevant, contextually appropriate, and authentic narrations was assessed.A qualitative survey is also planned to gather feedback from Islamic scholars, university students, and general users. Metrics such as perceived relevance, interpretability, and ease of use are used to refine the framework. This phase was essential in ensuring the practical utility and user acceptance of the proposed semantic search system.

4. Conceptual Framework

4.1 Overview

The framework integrates authenticated Hadith sources, semantic annotations, and NLP to enable context-aware retrieval. It supports multilingual queries (Arabic, English, Urdu) and is designed as a modular, theoretical blueprint for future implementation.

4.2 Layered Architecture

1. **Data Ingestion and Preprocessing:** Normalizes multilingual corpora, segments isnad and matn, and extracts metadata using tools like Farasa.
2. **Semantic Annotation and Enrichment:** Tags texts with themes, tone, and structure, combining manual and automated methods.
3. **Ontology and Knowledge Graph:** Uses OWL to map concepts and relationships, supporting reasoning and query expansion.
4. **Semantic Embedding and Retrieval:** Encodes texts with SBERT/AraBERT, indexes with FAISS, and retrieves based on semantic

similarity.

5. **User Interface and Interaction:** Offers natural language queries, faceted filtering, and multilingual result display.

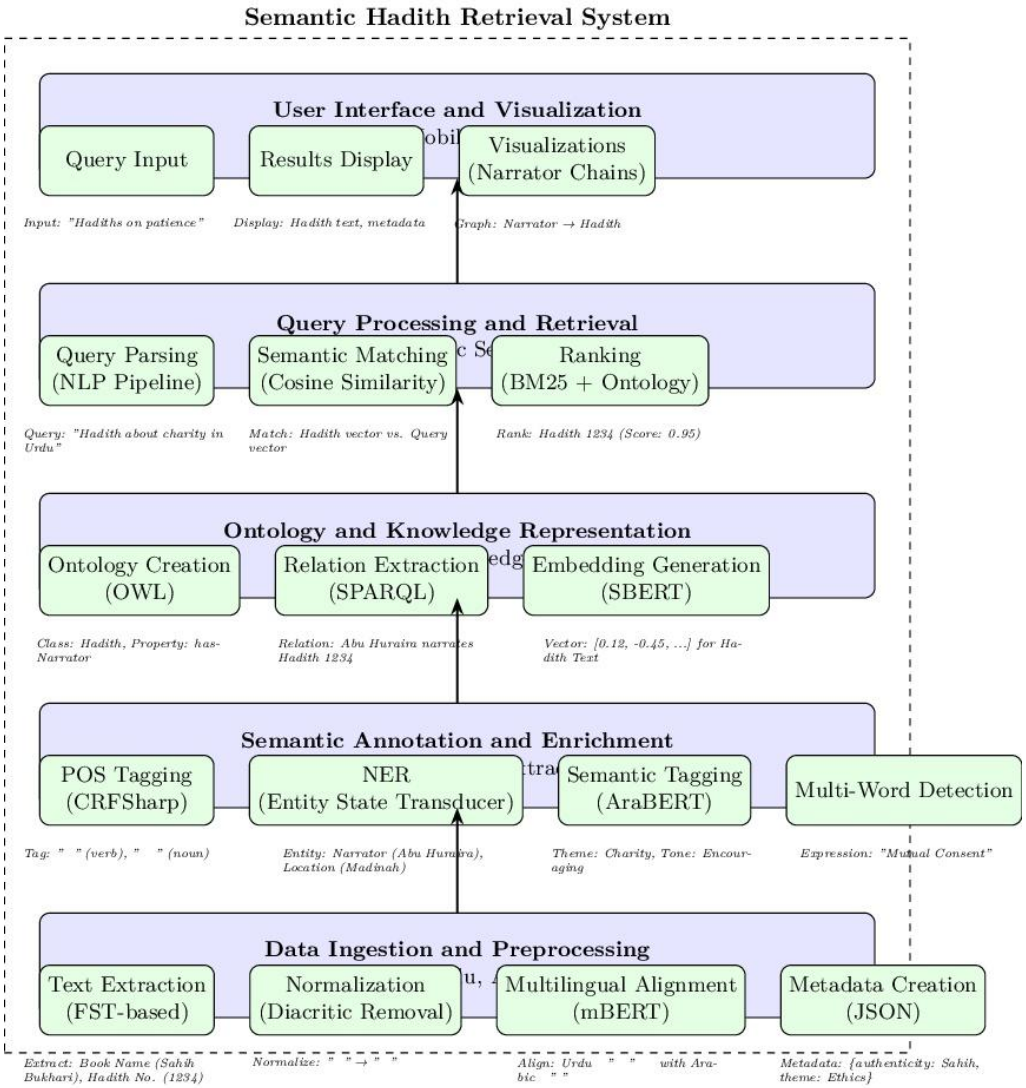

Figure 4: Layered Architecture of Conceptual Framework

4.3 Integration of NLP and Ontology

NLP tools (e.g., Farasa, AraBERT) extract linguistic features, while ontologies provide structured knowledge. This synergy enables conceptual matching, e.g., linking "mercy" to "compassion" or "jihad" to "warfare," enhancing retrieval accuracy.

4.4 Sample Query Processing: "Mercy in Warfare"

The table below illustrates how the query "Mercy in warfare" (Arabic: الرحمة في الحرب, Urdu: جنگ میں رحم) is processed, retrieving Sahih Bukhari Hadith #3015[50].

Layer	Process	English	Arabic	Urdu	Outcome
1: Data Ingestion & Preprocessing	Tokenize, normalize, segment, metadata extraction	Query: "Mercy in warfare" → ["mercy", "warfare"] Hadith #3015: "Do not kill women..."	Query: الرحمة في الحرب → [الرحمة, الحرب] Hadith: لا تقتلوا النساء...	Query: جنگ میں رحم → [جنگ, رحم] Hadith: عورتوں، بچوں کو نہ مارو	Cleaned corpus, metadata: Book (Sahih Bukhari), Hadith #3015
2: Semantic Annotation & Enrichment	Tag themes, tone, structure	Tags: ethics, warfare, compassionate Query tags: ethics, warfare	Tags: أخلاق، حرب، رحمة Query tags: same	Tags: اخلاق، جنگ، رحم Query tags: same	Hadith and query tagged with "ethics", "warfare"
3: Ontology & Knowledge Graph	Map to ontology concepts, relationships	Concepts: mercy → ethics, warfare → jihad Relation: hasTheme	Concepts: الرحمة → أخلاق, الحرب → جهاد Relation: له موضوع	Concepts: رحم → اخلاق, جنگ → جهاد Relation: موضوع ہے	Query mapped to "ethics", "jihad" nodes
4: Semantic Embedding & Retrieval	Encode, index (FAISS), match (cosine similarity)	Query embedding: SBERT("Mercy in warfare") Hadith #3015 embedding	Query embedding: AraBERT(الرحمة في الحرب) Hadith embedding	Query embedding: SBERT(جنگ میں رحم) Hadith embedding	Top result: Hadith #3015, score ~0.89
5: User Interface & Interaction	Display, highlight, filter	Display: "Do not kill women..." Highlight: "non-combatants"	Display: لا تقتلوا النساء... Highlight: غير المقاتلين	Display: عورتوں، بچوں کو نہ مارو Highlight: غیر جنگجو	Hadith #3015 shown with metadata, filters

Table 1: Sample Query Processing: "Mercy in Warfare

4.5 Technology Stack

The proposed Semantic Hadith Retrieval System leverages a robust technology stack to transform theoretical concepts into a practical, multilingual platform for accessing Hadith literature. Python drives the backend with libraries like spaCy, NLTK, and Hugging Face Transformers for NLP tasks, while React.js or Vue.js powers a responsive, multilingual

frontend. The system relies on a curated Hadith corpus from sources like Sunnah.com, stored in MongoDB or PostgreSQL with Elasticsearch for fast retrieval. Ontologies, built with OWL and Protégé, enable semantic reasoning, stored in RDF triple stores like Apache Jena and enhanced by external resources like Arabic WordNet. Protégé has extended support for plugins and has interoperability with other tools like Jena, HTML as well as UTF-8 support[51]. Transformer models (AraBERT, SBERT) generate dense vector embeddings, indexed via FAISS for efficient similarity searches, with re-ranking based on authenticity and themes. Evaluation uses intrinsic metrics (precision, recall) and extrinsic user feedback, tracked with MLflow and TensorBoard, providing a scalable, scholarly-aligned blueprint for developers.

5. Practical Considerations

5.1 Use Case Scenarios

To demonstrate the practicality and scope of the proposed conceptual framework, several hypothetical use case scenarios are presented. These scenarios illustrate how the system would function in real-world settings, assuming full implementation. They also help in visualizing the knowledge flow from user query to semantic retrieval.

Use Case 1: Scholar Searching for Legal Precedents A fiqh scholar is researching rulings related to "interest" (riba) in financial transactions. The scholar types a query in Arabic: "ما ورد عن الربا في المعاملات." The NLP module tokenizes and parses the query, identifies "riba" as a concept, and matches it with the ontology's legal domain node under "Muamalat." The embedding module retrieves Hadiths that are semantically similar, even if they don't contain the word "riba" explicitly but address concepts like "usury," "loan exploitation," or "unjust profit." The scholar receives them ranked by authenticity and relevance.

Use Case 2: Student Seeking Ethical Guidance A university student types in English: "What does Islam say about kindness to parents?" The system performs semantic parsing and links "kindness" and "parents" to ethical themes in the ontology. It retrieves Hadiths that discuss honoring parents, obligations of children, and Prophet Muhammad's sayings on familial respect. English translations are displayed with a toggle to view original Arabic. Contextual highlights pinpoint relevant passages.

Use Case 3: General Public User Exploring Emotional Topics A layperson, feeling anxious, searches: "Hadiths about managing anger." The system interprets "anger" as a psychological-emotional theme, linked in the ontology

under "Akhlaq." NLP tools detect the tone and intent, suggesting Hadiths with calming advice, recommended actions, and the Prophet's own practice of self-restraint. Related themes such as "patience" and "mercy" are suggested through semantic expansion.

In all these cases, the knowledge flow begins with the user query, moves through NLP parsing, semantic embedding, and ontological reasoning, and ends with a ranked, contextually relevant list of Hadiths. This end-to-end flow emphasizes how the conceptual framework bridges raw user input and deep Islamic knowledge, offering meaningful and authentic guidance.

Use Case 4: Scholarly Research and Thematic Compilation Islamic scholars and researchers often need to compile hadiths around specific topics such as ethics, governance, economic justice, or women's rights. Traditional methods require sifting through printed volumes or static digital archives, demanding substantial time and prior knowledge of Arabic lexicon. In the proposed semantic framework, scholars can perform thematic searches like "financial integrity in trade" or "accountability in leadership," retrieving narrations that are semantically relevant even if they do not use the exact queried terms.

For instance, a researcher looking into justice (ʿadl) may discover narrations not only explicitly mentioning justice but also those concerning fairness, honesty, or equitable treatment, owing to the system's ontological reasoning and semantic embeddings. The ability to trace narrations thematically across multiple canonical sources (e.g., Sahih Bukhari, Sahih Muslim, Sunan Abu Dawood) makes comparative analysis easier and more accurate.

Use Case 5: Religious Education and Curriculum Development Teachers in Islamic schools or seminaries can use the system to curate lesson plans centered around specific moral or theological themes. For example, if an instructor is preparing a class on "compassion in the Sunnah," the system can provide a filtered, curated list of Hadiths aligned with that concept, including those that use variant terms or are presented through parables and actions of the Prophet ﷺ rather than direct verbal statements. The semantic layer allows educators to break out of rigid keyword constraints and focus instead on meaning and context. Furthermore, the platform's user interface can support visual teaching aids such as concept maps, timelines of prophetic sayings, and thematic clusters, making it a powerful tool for classroom engagement.

Use Case 6: Personal Study and Ethical Guidance Everyday users, including students and laypersons, often search Hadith literature to find ethical or spiritual guidance. A person looking to understand Islamic teachings on anger management might enter a natural language query like "how did the Prophet

deal with anger?" Traditional systems might return irrelevant results or miss key narrations entirely. The semantic retrieval engine, however, would return Hadiths discussing anger directly, related incidents where the Prophet ﷺ showed patience, or stories illustrating emotional self-control—enriching the seeker's understanding and experience. In this way, the system not only increases accessibility but also supports the individual's spiritual development through deeper, contextual exploration of the Prophet's teachings.

Use Case 7: Legal and Fatwa Committees Fatwa councils and Islamic legal experts frequently rely on authentic Hadiths to support legal opinions (fatāwā). With semantic tagging, the system can assist muftis and jurists by providing conceptually aligned Hadiths related to modern problems. A query like "digital financial fraud" may retrieve Hadiths discussing fraudulent trade practices or deceptive contracts—even if the original text doesn't mention anything digital—thanks to ontology-based expansion. This capability is critical in ijtihād (independent reasoning), where analogical reasoning must bridge classical principles with contemporary issues. The framework thus becomes a powerful aid in dynamic jurisprudence.

Use Case 8: Cross-Language Accessibility for Global Users Given the multilingual nature of the Muslim ummah, English-speaking or non-Arabic users often struggle with accessing Hadiths due to translation inconsistencies. A semantic system that maps equivalent terms and paraphrases ensures that a user querying in English still receives narrations originally recorded in Arabic. This is made possible by a semantic mapping layer and cross-lingual embeddings that account for synonymous structures in translation. For instance, if one user searches "purity" and another "cleanliness," both would receive a consolidated result set, even if those terms appear differently in translated Hadiths. This significantly broadens access to non-specialist audiences without compromising on the scholarly integrity of the information.

5.2 Semantic Annotation Example

To demonstrate the operational depth and scholarly precision of the proposed semantic Hadith retrieval system, it is essential to explore how semantic annotation functions in practice. Semantic annotation refers to the process of enriching textual data with structured metadata, enabling the system to understand not just the words but their meanings, roles, and contextual implications. This section provides illustrative examples of semantic annotation applied to Hadith texts, showing how machine learning and linguistic tools are employed to extract and label nuanced information within

the narrations.

Consider the following Hadith from Sahih al-Bukhari:

"*The Prophet ﷺ said, 'The strong is not the one who overcomes the people by his strength, but the strong is the one who controls himself while in anger.'*"[52]

In a traditional retrieval system, only users who search for the exact phrase "strength" or "anger" may be able to find this narration. However, semantic annotation transforms this static text into a dynamic knowledge object. The annotation pipeline begins with syntactic parsing using Arabic NLP tools like Farasa[53] or CAMeL, breaking the sentence into parts of speech and identifying clauses. Named Entity Recognition (NER) detects the speaker (the Prophet ﷺ) and the contextual subject (anger management). Subsequently, dependency parsing identifies the relational structure between strength, control, and anger.

At the semantic level, the system assigns role labels to different components: the speaker (Prophet ﷺ) is tagged as the *authority*, the action ("said") as the *speech act*, and the clause "controls himself while in anger" is classified under the thematic ontology of *emotional self-regulation* and *character refinement*. The annotation also captures cross-references to Qur'anic concepts such as *ṣabr* (patience) and *taqwā* (God-consciousness), which are semantically related but not explicitly stated.

Furthermore, multilingual annotation ensures this Hadith is semantically aligned with equivalent English translations. Whether the word used is "anger," "rage," or "temper," they are linked to the same concept node in the ontology. This alignment is achieved through vectorized embeddings created using models like AraBERT [54] and multilingual Sentence-BERT, which represent both Arabic and English phrases in a shared semantic space.

Another example is a Hadith concerning trade ethics:

"*The Prophet ﷺ said: 'The buyer and the seller have the option (of canceling or confirming the bargain) as long as they have not separated...'" (Sahih Bukhari)*[55].

Here, semantic annotation first identifies the domain (Islamic finance), the role entities (buyer, seller), and the condition (physical separation). The action clause is annotated under *commercial contract law* and linked to ontological themes like *mutual consent*, *business ethics*, and *Islamic jurisprudence (fiqh al-muʿāmalāt)*.

The semantic enrichment also involves tagging the Hadith's narrator chain (isnād) and matn (main text) separately. The isnād is analyzed using named entity recognition to detect individual narrators and establish connections within the narrator network, which is stored in a graph database. This allows the system to support isnād-based filtering, where users may search for Hadiths narrated by or through specific individuals, or check the authenticity grade.

Ontologies used in the system—such as a domain-specific Islamic Ontology and WordNet synsets—also support intertextual annotation. For instance, when annotating Hadiths on *cleanliness*, related concepts like *ritual purity, ablution (wuduʾ),* and *physical hygiene* are all semantically linked. This enriches the retrieval process, enabling the user to search with broader conceptual intent rather than relying on keyword precision.

In a technical context, semantic annotation is stored as a set of RDF triples or JSON-LD structures. For example:

```
{
  "subject": "The Prophet",
  "predicate": "promotes",
  "object": "emotional self-control",
  "theme": "anger management",
  "authenticity": "Sahih",
  "narrator": ["Abu Huraira"],
  "text_language": "Arabic",
  "semantic_class": "Behavioral Ethics"
}
```

Such structures allow the retrieval system to index and reason over Hadith data in a machine-interpretable format. These annotations also feed into the re-ranking algorithms, giving preference to Hadiths with higher thematic relevance, broader contextual depth, or greater authenticity.

Ultimately, semantic annotation transforms static Hadith texts into dynamic, queryable knowledge graphs. It bridges the gap between unstructured religious texts and the analytical needs of contemporary users—be they scholars, educators, or laypersons. This transformation underlies the power of semantic retrieval and positions the system as a tool not only for access but for deep Islamic knowledge discovery.

5.3 Illustrative Evaluation Plan: Projected Metrics Across Development Phases

To estimate the potential performance of the proposed semantic Hadith

retrieval system, a set of projected evaluation metrics is outlined below. These metrics are based on established performance benchmarks of similar semantic retrieval models in other domains (e.g., biomedical, legal, and Quranic NLP systems), adapted to the Hadith context. The projections span two development phases:

- Phase I (Prototype Stage): Initial model with basic semantic embedding and ontology support.
- Phase II (Beta Stage): Model refined with user feedback, additional annotation, and optimized indexing.

The following table summarizes these illustrative projections across key evaluation categories:

Metric/Category	Method	Phase I: Projected	Phase II: Projected	Notes/Trends
Precision	% of relevant Hadiths retrieved	0.85 (85%)	0.88 (88%)	Improved through fine-tuned semantic embeddings
Recall	% of relevant Hadiths identified	0.82 (82%)	0.86 (86%)	Ontology expansion increases thematic coverage
F1-Score	Harmonic mean of Precision and Recall	0.83 (83%)	0.87 (87%)	Better balance of precision and recall expected
Mean Reciprocal Rank (MRR)	Ranking quality (avg. reciprocal rank)	0.79 (79%)	0.84 (84%)	Ranking enhanced by contextual embeddings
nDCG	Top-k ranking relevance	0.81 (81%)	0.85 (85%)	Top results become more relevant
Semantic Similarity Accuracy	Paraphrase detection (annotated pairs)	0.87 (87%)	0.90 (90%)	Improved embedding training
Ontology Reasoning Tests	SPARQL query success rate	0.80 (80%)	0.83 (83%)	Better inferencing over theme relations
User Surveys	Satisfaction score (1–5 scale)	4.2 (84%)	4.5 (90%)	UI/UX and retrieval interpretability refinements
Expert Validation Panels	Scholarly rating of semantic accuracy	4.3 (86%)	4.6 (92%)	Alignment with Hadith scholarship improves
Case Study Evaluation	% of tasks successfully completed	78%	85%	Real-world use cases show increasing effectiveness
Latency	Average response time (ms)	450 ms	320 ms	Optimized backend search and caching
Throughput	Queries processed per second (QPS)	50 QPS	75 QPS	Projected scalability improvements
Click-Through Rate (CTR)	% of users clicking retrieved results	70%	82%	More relevant and ranked results
Average Session Duration	Time spent per session	4.5 minutes	6.2 minutes	Increased engagement due to content richness
Failed Search Attempts	% of queries with no meaningful results	12%	7%	Query suggestions reduce null returns

Table 2: Projected Evaluation Metrics

6. Conclusion and Future Directions

This study presents a Semantic Retrieval System for Hadith literature, specifically focusing on Sahih Bukhari, by integrating advanced Natural Language Processing (NLP) techniques and ontology-based knowledge representation. The objective was to improve upon traditional keyword-based retrieval methods by introducing a semantically enriched framework that captures the contextual and thematic depth of Hadith content. Key components of the system include an Arabic NLP pipeline, a custom-built ontology for thematic and relational structuring, and a multilingual user interface. The system architecture was designed to support diverse user needs—ranging from scholars and educators to general audiences—by enabling nuanced search and exploration of Hadith texts. Named Entity Recognition (NER), semantic annotation, and ontological reasoning were employed to enhance precision and relevance in retrieval tasks. The research also outlined a detailed evaluation strategy covering performance, usability, and scholarly alignment. However, several challenges remain, including linguistic complexity in classical Arabic, translation inconsistencies, ongoing ontology management, and the need to accommodate diverse interpretive traditions within Islam.

Future work should focus on implementing a working prototype that combines Arabic NLP, ontology modeling, semantic search, and multilingual support. Adding Tafsir and Fiqh texts can improve context, while support for languages like Urdu, Turkish, and Persian will expand reach. Features such as quizzes, gamified learning, and mobile apps can boost engagement. A collaborative annotation tool will allow scholars and users to refine content. Integration with platforms like Al-Shamela via APIs and the use of explainable AI will enhance interoperability, transparency, and reliability.

Notes and References:

[1] Brown, J. A. (2017). Hadith: Muhammad's legacy in the medieval and modern world. Simon and Schuster.
[2] Kamali, M. H. (2014). A textbook of Hadith studies: authenticity, compilation, classification and criticism of Hadith. Kube Publishing Ltd.
[3] Dutton, Y. (2013). The Origins of Islamic Law: The Qur'an, the Muwatta'and Madinan Amal. Routledge.
[4] Hallaq, W. B. (1997). A history of Islamic legal theories: An introduction to Sunni Usul al-Fiqh.

Cambridge University Press.
[5] Nur'aini, L. H. (2025). The use of digital technology in hadith studies. At Turots: Jurnal Pendidikan Islam, 12-23.
[6] Alghamdi, M., Abushawarib, M., Ellouh, M., Ghaleb, M., & Felemban, M. (2023, December). Enhancing arabic information retrieval for question answering. In Proceedings of the 7th International Conference on Future Networks and Distributed Systems (pp. 366-371).
[7] Alowaidi, S., Atwel, E., & Alsalka, M. A. (2024). Survey of Semantic Islamic Search Systems. International Journal on Islamic Applications in Computer Science And Technology, 12(4).
[8] Daud, A., Ullah, M. H., Banjar, A. R., & Alshdadi, A. A. (2022). Ontological modeling and semantic search in quran. IJCSNS, 22(5), 771.
[9] Azmi, A. M., Alkhalifah, F., Alsaeed, A., & Barnawi, Y. (2017, September). Using non-conventional search schemes to retrieve Hadiths. In The 5th international conference on Arabic language processing (CITALA'14), Oujda, Morocco. http://www. citala. org/citala2014/papers/paper_39. pdf. Accessed (Vol. 11).
[10] https://shamela.ws/
[11] Ibda, H., Sofanudin, A., Syafi, M., Soedjiwo, N. A. F., Azizah, A. S., & Arif, M. (2023). Digital learning using Maktabah Syumilah NU 1.0 software and computer application for Islamic moderation in pesantren. International Journal of Electrical and Computer Engineering, 13(3), 3530-3539.
[12] Li, H., & Xu, J. (2014). Semantic matching in search. Foundations and Trends® in Information Retrieval, 7(5), 343-469.
[13] Tamine, L., & Goeuriot, L. (2021). Semantic information retrieval on medical texts: Research challenges, survey, and open issues. ACM Computing Surveys (CSUR), 54(7), 1-38.
[14] Xiong, C., Power, R., & Callan, J. (2017, April). Explicit semantic ranking for academic search via knowledge graph embedding. In Proceedings of the 26th international conference on world wide web (pp. 1271-1279).
[15] Castells, P., Fernandez, M., & Vallet, D. (2006). An adaptation of the vector-space model for ontology-based information retrieval. IEEE transactions on knowledge and data engineering, 19(2), 261-272.
[16] Kim, W. J., Kim, D. H., & Jang, H. W. (2016). Semantic extention search for documents using the Word2vec. The Journal of the Korea Contents Association, 16(10), 687-692.
[17] Zhang, L. (2025). Improved Web Page Categorization with Semantic-Aware Focused Crawling Using GloVe and TF-IDF. J. COMBIN. MATH. COMBIN. COMPUT, 127, 6569-6586.
[18] Trisnawati, L., Samsudin, N. A. B., Bin Ahmad Khalid, S. K., Bin Ahmad Shaubari, E. F., & Indra, Z. (2025). An Ensemble Semantic Text Representation with Ontology and Query Expansion for Enhanced Indonesian Quranic Information Retrieval. International Journal of Advanced Computer Science & Applications, 16(1).
[19] Al-Sanasleh, H. A., & Hammo, B. H. (2017, October). Building domain ontology: Experiences in developing the prophetic ontology form Quran and hadith. In 2017 International Conference on New Trends in Computing Sciences (ICTCS) (pp. 223-228). IEEE.
[20] AlZahrani, F. M., & Al-Yahya, M. (2023). A transformer-based approach to authorship attribution in classical arabic texts. Applied Sciences, 13(12), 7255.
[21] Almutrash, S., & Abudalfa, S. (2024). Comparative Study on the Efficiency of Using PaLM and CAMeLBERT for Arabic Entity Sentiment Classification.
[22] Sellami, M., Hadrouk, R., Chelghoum, S., Badache, R., Kamel, N., & Lakhfif, A. (2024, November). Multitask Fake News Detection in Arabic Language using AraELECTRA model:

COVID-19 Case Study. In 2024 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM) (pp. 1-7). IEEE.

[23] Nazri, N. A. B. M., & Omar, A. W. B. (2025). Fine-tuning Large Language Model (BERT) for Islamic Moral Inquiry and Response. International Journal on Perceptive and Cognitive Computing, 11(1), 88-94.

[24] Fairouz, B., Nora, T., & Nouha, A. A. (2020). An ontological model of hadith texts. International Journal of Advanced Computer Science and Applications, 11(4), 2020.

[25] Jaafar, A. H., & Che Pa, N. (2016). Hadith commentary repository: An ontological approach.

[26] Alkhatib, M., Monem, A. A., & Shaalan, K. (2017). A Rich Arabic WordNet Resource for Al-Hadith Al-Shareef. Procedia Computer Science, 117, 101-110.

[27] Harrag, F. (2014). Text mining approach for knowledge extraction in Sahîh Al-Bukhari. Computers in Human Behavior, 30, 558-566.

[28] Al-Arfaj, A., & Al-Salman, A. (2014, September). Towards ontology construction from Arabic texts-a proposed framework. In 2014 IEEE International Conference on Computer and Information Technology (pp. 737-742). IEEE.

[29] Azmi, A. M., Al-Qabbany, A. O., & Hussain, A. (2019). Computational and natural language processing based studies of hadith literature: a survey. Artificial Intelligence Review, 52(2), 1369-1414.

[30] Aldhlan, K. A., Zeki, A. M., & Zeki, A. M. (2010, December). Datamining and Islamic knowledge extraction: alhadith as a knowledge resource. In Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010 (pp. H-21). IEEE.

[31] Naji Al-Kabi, M., Kanaan, G., Al-Shalabi, R., Al-Sinjilawi, S. I., & Al-Mustafa, R. S. (2005). Al-Hadith text classifier. Journal of Applied Sciences, 5(3), 584-587.

[32] Saeed, S., Yousuf, S., Khan, F., & Rajput, Q. (2022). Social network analysis of Hadith narrators. Journal of King Saud University-Computer and Information Sciences, 34(6), 3766-3774.

[33] Azmi, A., & Badia, N. B. (2010, August). iTree-Automating the construction of the narration tree of Hadiths (Prophetic Traditions). In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010) (pp. 1-7). IEEE.

[34] Azmi, A., & Badia, N. B. (2010, August). iTree-Automating the construction of the narration tree of Hadiths (Prophetic Traditions). In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010) (pp. 1-7). IEEE.

[35] Baraka, R. S., & Dalloul, Y. (2014). Building Hadith ontology to support the authenticity of Isnad. International Journal on Islamic Applications in Computer Science And Technology, 2(1), 25-39.

[36] Altammami, S., Atwell, E., & Alsalka, A. (2020). The Arabic-English parallel corpus of authentic hadith. International Journal on Islamic Applications in Computer Science And Technology, 8(2), 1-10.

[37]https://sunnah.com/

[38] http://www.shamela.ws/

[39]Darwish, K., & Mubarak, H. (2016, May). Farasa: A new fast and accurate Arabic word segmenter. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1070-1074).

[40]Darwish, K., & Mubarak, H. (2016, May). Farasa: A new fast and accurate Arabic word segmenter. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1070-1074).

[41] Alfaidi, A., Alwadei, H., Alshutayri, A., & Alahdal, S. (2023). Exploring the performance of farasa and CAMeL taggers for arabic dialect tweets. Int. Arab J. Inf. Technol., 20(3), 349-356.
[42] Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

[43]https://protege.stanford.edu/

[44]Alatrish, E. S. (2013). Comparison some of ontology. Journal of Management Information Systems, 8(2), 018-024.
[45] Dahir, S., Khalifi, H., & El Qadi, A. (2019, March). Query expansion using DBpedia and WordNet. In Proceedings of the ArabWIC 6th Annual International Conference Research Track (pp. 1-6).
[46] Mutawa, A. M., & Sruthi, S. (2025). A Comparative Evaluation of Transformers and Deep Learning Models for Arabic Meter Classification. Applied Sciences, 15(9), 4941.
[47] Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2020). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. arXiv preprint arXiv:2010.12858.
[48] Sibaee, S., Ahmad, S., Khurfan, I., Sabeeh, V., Bahaaulddin, A., Belhaj, H., & Alharbi, A. (2023, December). Qamosy at Arabic reverse dictionary shared task: Semi decoder architecture for reverse dictionary with SBERT encoder. In Proceedings of ArabicNLP 2023 (pp. 467-471).
[49] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., ... & Jégou, H. (2024). The faiss library. arXiv preprint arXiv:2401.08281.
[50]Sahih al-Bukhari 3015 https://sunnah.com/bukhari:3015
[51] Kamran, A. B., Abro, B., & Basharat, A. (2023). SemanticHadith: An ontology-driven knowledge graph for the hadith corpus. Journal of Web Semantics, 78, 100797.
[52]Sahih al-Bukhari 6114 https://sunnah.com/bukhari:6114
[53]Hammouda, T., Jarrar, M., & Khalilia, M. (2024). SinaTools: Open Source Toolkit for Arabic Natural Language Processing. Procedia Computer Science, 244, 388-396.
[54]Abo-Elghit, A. H., Hamza, T., & Al-Zoghby, A. (2022). Embedding Extraction for Arabic Text Using the AraBERT Model. Computers, Materials & Continua, 72(1).
[55]Sahih al-Bukhari 2109 https://sunnah.com/bukhari:2109