Jawahir Al-Ahadis
e-ISSN: 3110-0597
Vol. 1 No. 2 Edisi Desember 2025

ORIGINAL ARTICLE OPEN ACCESS

# Digitalisasi dan Rekonstruksi Struktural al-Mujam al-Mufahras: Sebuah Model Integrasi Leksikon Tradisional ke dalam Repositori Hadis Digital Terintegrasi

**Nandi Pinto[1*], Siska Julianty[2], Anton Hilman[3]**

***Korespondensi:**
email:
nandipintoamrizal@uinib.ac.id
[1,2] Universitas Islam Negeri Imam Bonjol Padang, Indonesia
[3] Profesional Programmer PT. Dimensitechno, Indonesia

Sejarah Artikel:
Submit: 15 Juli 2025
Revisi: 10 Oktober 2025
Diterima: 15 November 2025
Diterbitkan: 29 Desember 2025

**Kata Kunci:**
Repositori Hadis Digital, Integrasi Data Leksikal, Pemodelan Entitas-Hubungan, Pencocokan Semantik, Filologi Digital, (NLP)

**Abstrak**

Revolusi digital dalam preservasi naskah klasik masih menyisakan celah dalam pendigitalan karya leksikal kompleks seperti al-Mujam al-Mufahras li Alfaz al-Hadith al-Nabawi. Padahal, karya ini merupakan *concordance* yang esensial untuk akses semantik terhadap korpus hadis. Pertanyaan penelitian utama yang diangkat adalah: Bagaimana merancang model integrasi data yang mampu mendigitalkan struktur analog al-Mujam al-Mufahras ke dalam repositori hadis digital yang terstandarisasi dan semantik? Menggunakan pendekatan Research and Development (R&D), penelitian ini melakukan dekonstruksi struktural terhadap 1.200 entri sampel, lalu merancang Entity-Relationship Model (ERM) yang menghubungkan *lemma*, kutipan kontekstual, dan referensi silang dengan teks lengkap *Kutub al-Tisah*. Implementasi dilakukan melalui algoritma hibrida: Reference Matching (untuk referensi eksplisit) dan Semantic Matching berbasis NLP (cosine similarity) untuk kasus ambigu. Hasil pengujian menunjukkan *F1-Score* 90%, dengan validasi pakar mencapai 94%. Model ini membuktikan bahwa digitalisasi leksikon hadis bukan hanya alih media, melainkan rekonstruksi pengetahuan yang memungkinkan penemuan ilmiah berbasis semantik dalam studi hadis, filologi digital, dan linguistik Arab.

***Abstract***

*The digital revolution in the preservation of classical manuscripts still leaves a gap in the digitization of complex lexical works such as al-Mujam al-Mufahras li Alfaz al-Hadith al-Nabawi. In fact, this work is an essential concordance for semantic access to the hadith corpus. The main research question raised is: How to design a data integration model capable of digitizing the analogue structure of al-Mujam al-Mufahras into a standardized and semantic digital hadith repository? Using a Research and Development (R&D) approach, this study conducted a structural deconstruction of 1,200 sample entries, then designed an Entity-Relationship Model (ERM) that connects lemmas, contextual citations, and cross-references with the full text of the Polar al-Tis'ah. Implementation is carried out through a hybrid algorithm: Reference Matching (for explicit references) and NLP-based Semantic Matching (cosine similarity) for ambiguous cases. The test results showed an F1-Score of 90%, with expert validation reaching 94%. This model proves that the digitization of the hadith lexicon is not only a media transfer, but a reconstruction of knowledge that allows semantics-based scientific discoveries in the study of hadith, digital philology, and Arabic linguistics.*

## 1. PENDAHULUAN

Revolusi digital telah membawa dampak transformatif pada preservasi khazanah keilmuan klasik, termasuk dalam bidang ilmu hadis (Iryani et al., 2023). Inisiatif digitalisasi besar-besaran telah dilakukan terhadap berbagai kitab induk (*Kutub al-Tisah*) seperti Shahih al-Bukhari dan Shahih Muslim, yang umumnya berfokus pada konversi narasi teks lengkap (*matan*) dan pembangunan basis data rantai periwayatan (*sanad*) ke dalam format yang dapat dibaca mesin (Aziz et al., 2022). Namun, akses leksikal-semantik terhadap korpus hadis tersebut masih sangat terbatas. Karya indeks pra-digital sepertikitab al-Mujam al-Mufahras li Alfaz al-Hadith al-Nabawi (selanjutnya ditulis al-Mujam al-Mufahras). Padahal, kitab ini merupakan indeks leksikal paling komprehensif dari abad kedua hijriah, yang tidak hanya merekam keberadaan kata dalam hadis, tetapi juga menyediakan sistem navigasi berbasis kode

simbolis (misalnya, simbol خ untuk merujuk Shahih al-Bukhari) dan format penomoran multi-lapis (seperti 89 : خ النكاح) sebagai bentuk kompresi data dan representasi semantik yang canggih di era analog (Baalbaki, 2014; Hitti et al., 1936). Karena itu, penelitian ini berargumen bahwa digitalisasi yang efektif harus bergeser dari paradigma alih media semata menuju paradigma rekonstruksi dan integrasi pengetahuan, yang memungkinkan pemanfaatan penuh fungsi al-Mujam al-Mufahras sebagai alat penemuan pengetahuan dalam ekosistem digital.

Penelitian terdahulu terfokus pada dua arus utama yakni: *Pertama,*repositori teks lengkap seperti SemanticHadith oleh (Kamran et al., 2023), Shamela oleh (Wahid & Wahyuni, 2018), Jawami al-Kalim dan Lidwa Pustaka oleh (Fauzi, 2020), dan dorar.net oleh (Najiyah & Putriani, 2024). *Kedua,* Analisis Jaringan Sanad (Social Network Analysis) yang disampaikan oleh (Dalimunthe & Siti, 2021; Guellil et al., 2021; Pinto et al., 2022; Saeed et al., 2022; Suhendra, 2019; Wahyuningsih & Istianah, n.d.; Yeni et al., 2024). Namun, tidak ada penelitian yang secara khusus merancang kerangka integrasi untuk karya indeks leksikal, yang memiliki struktur berbasis *lemma*, kutipan parsial, dan referensi silang multi-layer. Penelitian ini dilakukan untuk merespons celah metodologis yang diidentifikasi melalui penelitian terdahulu dalam digitalisasi naskah keagamaan yang masih didominasi oleh paradigma digital sebagai repositori dengan pemahaman teks dialihmediakan untuk tujuan preservasi dan pencarian kata kunci sederhana. Sebagaimana pencarian leksikal ini tidak mewakili dan memadai untuk menangani kompleksitas semantik dan struktural dari sebuah karya indeks seperti al-Mujam al-Mufahrash dengan sistem referensi silangnya.

Penelitian ini bertujuan menjawab: Bagaimana merancang model integrasi data untuk mendigitalkan struktur al-Mujam al-Mufahras secara utuh ke dalam repositori hadis digital terintegrasi? Secara operasional, penelitian ini (1) mendekonstruksi struktur data leksikon; (2) merancang ERM ternormalisasi; dan (3) mengimplementasikan algoritma pencocokan hibrida. Kemudian pertanyaan utama ini dijabarkan menjadi beberapa tujuan operasional sebagai berikut: *pertama*, untuk melakukan dekonstruksi struktural terhadap komponen-komponen data al-Mujam al-Mufahrash (*lemma*, cuplikan redaksi, sistem referensi silang) dan menganalisis pola kompleksitasnya. *Kedua*, untuk merancang sebuah Entity-Relationship Model (ERM) yang mengintegrasikan data leksikal dari al-Mujam al-Mufahrash dengan struktur data hadis lengkap (matan, sanad, dan metadata) dalam sebuah skema *database* yang ternormalisasi.

Kontribusi utama artikel ini adalah pergeseran paradigma digitalisasi dari *media conversion* menjadi *knowledge reconstruction*. Model ini memungkinkan leksikon klasik berfungsi sebagai lapisan semantik dinamis sehingga dapat diadopsi untuk karya leksikal Islam lainnya ketimbang memperlakukan al-Muam al-Mufahras sebagai dokumen statis yang cukup dipindai atau dikonversi ke format PDF atau teks digital, tetapi dirubah menjadi sistem pengetahuan dinamis yang memiliki logika internal, struktur relasional, dan fungsi epistemologis sebagai alat penemuan makna dalam korpus hadis. Dengan melakukan dekonstruksi struktural dan merancang model integrasi berbasis *Entity-Relationship* serta algoritma pencocokan hibrida, penelitian ini menghidupkan kembali fungsi asli leksikon tersebut dalam ekosistem digital, bukan hanya sebagai arsip mati, melainkan sebagai lapisan semantik aktif yang memungkinkan penelusuran kontekstual, koneksi antar-teks, dan penemuan pengetahuan berbasis data. Lebih jauh, arsitektur model yang diusulkan bersifat adaptif dan dapat direplikasi untuk karya leksikal Islam klasik lainnya seperti *Lisān al-'Arab, Tāj al-'Arūs, atau al-Mu'jam al-Wasīṭ* dan lain sebagainya sehingga membuka jalan bagi transformasi digital yang tidak hanya teknis, tetapi juga filologis, linguistik, dan epistemologis dalam studi Islam kontemporer.

## 2. KAJIAN PUSTAKA / TINJAUAN LITERATUR

### 2.1. *Digital Preservation dan Filologi Digital*

*Digital Preservation* mengacu pada serangkaian kegiatan dan strategi yang dilakukan untuk memastikan aksesibilitas dan keutuhan materi digital dalam jangka panjang (Brown, 2018). Dalam

konteks filologi digital, pendekatan ini tidak hanya sekadar mengalihmediakan naskah ke format digital, tetapi juga merekonstruksi struktur pengetahuan yang terkandung di dalamnya (Driscoll & Pierazzo, 2016). Pada kajian hadis, *Digital Preservation* telah diaplikasikan dalam pembangunan repositori teks lengkap dan basis data sanad, namun seringkali mengabaikan dimensi leksikal dan semantik yang menjadi inti dari karya indeks seperti al-Mu'jam al-Mufahrash. Studi oleh (Kamran et al., 2024) tentang *SemanticHadith* menunjukkan potensi grafik pengetahuan dalam mengorganisir relasi semantik, namun belum menyentuh aspek dekonstruksi struktural karya leksikon.

### 2.2. *Natural Language Processing* (NLP) untuk Teks Arab Klasik

*Natural Language Processing* (NLP) merupakan cabang ilmu komputer yang berfokus pada interaksi antara komputer dan bahasa manusia (Lehnert, 1992). Penerapan NLP pada teks Arab klasik, termasuk hadis, menghadapi tantangan khusus seperti variasi ortografi, morfologi yang kompleks, dan kurangnya data terannotasi (Guellil et al., 2021). Teknik seperti *tokenization*, *lemmatization*, dan *cosine similarity* telah digunakan untuk menangani ambiguitas dalam pencocokan teks. Namun, penerapannya pada karya indeks leksikal yang memiliki sistem referensi silang *multi-layer* masih belum banyak dieksplorasi.

### 2.3. Kerangka Konseptual / Posisi Penelitian

Penelitian ini memadukan teori *Digital Preservation* dan NLP untuk membangun sebuah kerangka integratif dalam mendigitalkan al-Mu'jam al-Mufahras. Pendekatan ini tidak hanya melihat digitalisasi sebagai preservasi konten, tetapi juga sebagai rekonstruksi pengetahuan yang memungkinkan kitab tersebut berfungsi sebagai alat penemuan yang dinamis. Gap teoritis yang diisi oleh penelitian ini adalah kurangnya model yang secara spesifik dirancang untuk mentransformasikan karya indeks leksikal analog menjadi lapisan semantik dalam repositori digital. Dengan menggabungkan dekonstruksi struktural dan pemodelan data relasional, penelitian ini menawarkan sudut pandang baru dalam integrasi leksikon tradisional ke dalam ekosistem digital.

## 3. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan Research and Development (R&D) dengan desain studi kasus, yang bertujuan menghasilkan model integrasi data berbasis rekonstruksi pengetahuan, bukan sekadar alih media. Desain ini didasarkan pada asumsi utama bahwa al-Mujam al-Mufahras walaupun berbentuk indeks analog memiliki struktur logis internal yang sistematis, terdiri atas *lemma* morfologis, kutipan kontekstual, dan sistem referensi silang multi-layer, yang secara prinsip dapat dipetakan ke dalam skema relasional digital tanpa kehilangan fungsi epistemologisnya. Logika desain penelitian dibangun di atas tiga langkah operasional utama *pertama,* Dekonstruksi struktural terhadap 1.200 entri sampel (halaman 1–75) untuk mengidentifikasi dan mendokumentasikan pola komponen data; *kedua,* Perancangan *Entity-Relationship Model* (ERM) yang ternormalisasi hingga *Boyce-Codd Normal Form* (BCNF), menghubungkan entitas *Lemma*, Kutipan_Kontekstual, Referensi_Silang, dan Hadis_Lengkap dalam satu kerangka terintegrasi; dan *Ketiga,* Implementasi algoritma hibrida yang menggabungkan *Reference Matching* (untuk referensi eksplisit) dan *Semantic Matching* berbasis NLP (*cosine similarity*) untuk menangani ambiguitas dan inkonsistensi referensi.

Keberhasilan model diukur melalui empat ukuran keberhasilan utama: *pertama,*Kinerja teknis sistem, dievaluasi menggunakan metrik precision, recall, dan *F1-Score* terhadap seluruh 1.200 entri. *kedua,* Akurasi integrasi end-to-end, diuji melalui integration testing pada 50 entri kompleks. *ketiga,* Validasi ahli, yang dilakukan oleh pakar ilmu hadis terhadap 100 entri (diperluas dari 50 entri awal untuk memperkuat validitas representatif), menghasilkan tingkat kecocokan. *keempat,* Kemampuan sistem untuk menjembatani *lemma* ke teks hadis lengkap secara semantic yang tidak hanya leksikal sehingga memungkinkan penelusuran kontekstual yang bermakna.

# 4. HASIL DAN PEMBAHASAN

## 4.1. Hasil Penelitian

### 4.1.1. *Dekonstruksi Struktur Al-Mu'jam* sebagai Karya Indeks

Proses ini dilakukan terahadap 1.200 entri sampel dari halaman 1–75 kitab al-Mujam al-Mufahras berhasil mengidentifikasi tiga komponen struktural inti ynag menjadi pilar penyusunan kamus ini, adapun ketiga komponen ini yakni

**a. Kata Kunci (Lemma)**

*Lemma*, atau kata kunci merupakan entri utama yang menjadi pintu masuk pencarian hadis. Hasil dekonstruksi menunjukkan bahwa penyusunan *lemma* tidak dilakukan secara acak, melainkan mengikuti sistem yang tertib dan logis. Susunan Abjad Arab disusun secara ketat berdasarkan urutan abjad hurufArab. Pola ini konsisten di seluruh sampel yang dianalisis, dimulai dari kata-kata berakar dari huruf ألف (*Alif*) dan همزة (*Hamzah*), seperti أبل ,أبر ,أبد, diikuti oleh إبل (dengan alif dan hamzah di bawah), إيثار, dan seterusnya mengikuti urutan hijaiyah. Sistem pengurutan ini, sebagaimana dijelaskan oleh Heywood dkk., dalam kamus Arabnya bahwa metode klasik yang digunakan memudahkan dalam pencarian manual terhadap kata, namun memerlukan normalisasi dalam lingkungan digital untuk menangani variasi penempatan hamzah(Cachia et al., 1985; Haywood et al., 1980).

Sampel susunan *database* dalam bentuk kata kunci/*lemma* sebagai berikut (data utuh terdapat dalam lampirkan):

**Tabel 1.** Struktur *Database Lemma*/Kata Kunci

| No | Abjad kata | Kosa kata | Redaksi hadis | Hasil temuan di kitab sumber | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ابد | أبدً | هذه البهائم لها أوابد كأوابدالوحش | خ | م | د | ت | ن | ق | دى | حل | ط | جه | Hlm | Jld | Thn |
| | | | | خهاد ١٩١ | أضا حى ٣٠ | أضا حى ١٤ | صيد ١٩ | صيد ٣٥،١٧ | ذباح ١٧،٩ | أضا حي ١٥ | (٣)٣٦٤،٤٦٣ | | | 1 | 1 | 1936 |
| | | | | شركه ٣, ١٦ | | | | ضحايا ٢٦ | | | | | | 1 | 1 | 1936 |
| | | | | ذباح ٣٧،٣٦،٢٣،١٨،١٥ | | | | | | | | | | 1 | 1 | 1936 |
| 2 | | أبدً | الله الابد | | | | | | دعاء ١٠ | | | | | 1 | 1 | 1936 |
| | | | هل لكم في الفلاح والرشد أ خر الابد | تفسير سورة ٤٠٣ | | | | | | | | | | 1 | 1 | 1936 |
| | | | ماكان وماهوكائن الى الابد | | | | قدر ١٧ | | | | | | | 1 | 1 | 1936 |
| | | | | | | | تفسير سورة ٦٨ | | | | | | | 1 | 1 | 1936 |
| | | | يا رسول الله ألعاملنا هذا أم لابد قال لابد | | حج ١٤١ | | | حج ٧٦ | مناسك ٨٣،٤١ | | ١٧٥(٣) | | | 1 | 1 | 1936 |
| | | | بل لأبد أبدا | | | | | | | مناسك ٣٤ | | | | 1 | 1 | 1936 |
| | | | ألنا هذه خاصة قال بل لأبد (و قرئ الأبد) | عمرة ٦ | | مناسك ٥٦،٢٣ | | | | | | | | 1 | 1 | 1936 |
| | | | | شركة ١٥ | | | | | | | | | | 1 | 1 | 1936 |
| | | | | تمنى ٣ | | | | | | | | | | 1 | 1 | 1936 |
| | | | سرية تخرج في سبيل الله أبدا | | | | | | جهاد ١ | | | | | 1 | 1 | 1936 |
| | | | اخرجه كما كنت أخرجه أبدا ما عشت | | زكاة ١٨ | | | | زكاة ٣١ | | | | | 1 | 1 | 1936 |
| | | | نحن الذين با يعون ا محمدا* على الجهاد ما بقينا أبدا (رجز) | مغازى ٢٩ | جهاد ١٣٠ | | | | | | | | | 1 | 1 | 1936 |
| | | | من شرب منه شربه لم يظمأ بعدها أبدا | | | | | | زهد ٣٦ | | | | | 1 | 1 | 1936 |
| | | | أن تحيوا فلا تمو توا أبدا | | | | | | | | ٢٨ (٣) | | | 1 | 1 | 1936 |
| | | | ان لكم أن تشبوا فلا تهوتوا أبدا | | | | | | | | ٢٨ (٣) | | | 1 | 1 | 1936 |
| | | | بالية فقلت لا تحيا أبدا | | | | | | | | ١٢ (٤) | | | 1 | 1 | 1936 |
| | | | لاأوبك الى ولا تحلين لى أبد | | | | | | | | | طلاق ٨٠ | | 1 | 1 | 1936 |
| | | | فوا الله لئن تغظنا رسول الله يمينه لا نفلح أبدا | ذبائح ٣٦ | | | | | | | | | | 1 | 1 | 1936 |
| | | | لا يجتمع في الدار كافر وقاتله أبدا | | | جهاد ١٠ | | | | | | | | 1 | 1 | 1936 |
| | | | ثم اعتدت من الاخر ثم لا يجتمعان أبدا | | | | | | | | | نكاح ٣٧ | | 1 | 1 | 1936 |
| | | | و من دخله لم يظمأ أبدا | | | | | | صيام ١ | | | | | 1 | 1 | 1936 |
| | | | لا صام من صام الابد | صوم ٥٧ | صيام ١٨٦، ١٨٧ | | | صيام ٧١،٧٨ | صيام ٢٨ | | | | | 1 | 1 | 1936 |
| 3 | ابزن | أبَرَ | وهم يأبرون النخل يقول يلقحون النخل | | فضائل ١٤٠ | | | | | | | | | 1 | 1 | 1936 |
| | | | له مهرة مأمورة أو سكة مأبورة | | | | | | | | ٣٦٨(٣) | | | 1 | 1 | 1936 |
| 4 | | أبرَ | أيما امرئ أبر نخلا | بيوع ٩٢ | بيوع ٧٩ | | | بيوع ٧٥ | | | | | | 1 | 1 | 1936 |
| | | | فا لتمر للذي أبر ها | بيوع ٩٢،٩٠ | بيوع ٧٩،٧٨،٧٦ | | | | تجارات ٣١ | | ٣٠،٥٣،٧٨،٥٢٦(٢) | | | 1 | 1 | 1936 |
| | | | من با ع نخلا قد أبرت فثمرتها للبائع | بيوع ٩٠ | | بيوع ٤٢ | | | تجا راة ٣١ | | ٢، ٩، ٢٣، ٧٨، ٨٢، ١٠٢، ١٥٠ [٢] | بيوع ٩ | | 1 | 1 | 1936 |
| | | | | مساقاة ١٧ | | | | | | | | | | 1 | 1 | 1936 |
| | | | | شروط ٢ | | | | | | | | | | 1 | 1 | 1936 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1936 | 1 | 1 | بيوع ٧٨ | | | | | | | | | أيما نخل اشترى أصو لها و قد أبرت | | |
| 1936 | 1 | 1 | | | | بيوع ٧٦ | بيوع ٣٥ | | بيوع ٨٠ | مساقاة ١٧ | | من با ع نخلا بعد أن تؤبر | | |
| 1936 | 1 | 1 | | [١٢٣]٦ | رهون ١٥ | | | | | | | لولم يفعلوا لصلح ظم يؤ بروا عا منذ | | |
| 1936 | 1 | 1 | | | | | بيوع ٢٥ | | | | | فى ابتيا ع النخل بعد التأبير | | |
| 1936 | 1 | 1 | | [٢٣٦،٣٦٥ ،٣٥٧]١ | | | | | | | ابا رُ | من السقى بعد الا با ر بشهرين | | 5 |

dst

*Sumber: Database* Hasil Penelitian, 2025

Berdasarkan sampel *database* di atas terlihat bentuk variasi dari masing-masing *lemma* atau kata kunci utama tidak berdiri sendiri, melainkan dilengkapi dengan berbagai bentuk variasi ortografis dan morfologisnya. Sebagai contoh, *lemma* أبد juga mencakup bentuk *adverbia* أبداً. Ini menunjukkan bahwa kamus ini tidak hanya mendaftarkan kata dasar, tetapi juga penggunaan praktisnya dalam bahasa, di mana perubahan bunyi dan penulisan sangatlah umum. Pendekatan ini merefleksikan pemahaman yang mendalam tentang fleksibilitas bahasa Arab. *Lemma* turunan dan derivasi morfologis lebih dari sekadar variasi, hal ini terlihat dari peta jejaring kata yang ditampilkan melalui proses derivasi.

Dari sebuah akar kata, kamus mencatat berbagai bentuk turunannya. Contoh yang teridentifikasi adalah dari akar أ-ب-ر, yang tidak hanya muncul sebagai أَبَرَ (bentuk kata kerja dasar), tetapi juga إِبَارَة (kata benda yang menunjukan alat atau tempat) dan مُؤَابَرَة (bentuk partisipial atau kata benda verbal yang mengandung makna resiprokal). Pola ini membuktikan bahwa al-Mujam al-Mufahras berfungsi tidak hanya sebagai indeks, tetapi juga sebagai sebuah *tesaurus* morfologis ringkas yang mengelompokkan kosakata berdasarkan akar semantiknya, sebuah fitur yang sangat berharga untuk analisis linguistik korpus hadis (Ryding, 2005).

**b. Kutipan Redaksi Hadis**

Setiap *lemma* dilengkapi dengan satu atau lebih kutipan (*shawāhid*) yang merupakan kutipan langsung dari potongan redaksi hadis Nabi sebagaimana data pada tabel 1 di atas. Kutipan ini berfungsi sebagai bukti penggunaan dan konteks di mana kata tersebut muncul. Fungsi kutipan ini adalah jantung dari kamus ini, yang memberikan "contoh hidup" dari penggunaan kata, jauh melampaui sekadar definisi leksikal. Misalnya, *lemma* أَوْابِد (bentuk jamak dari *wābidah* ynag berarti hewan liar) disertai dengan kutipan لَهُ أُوْابِدُ كَأَوَابِدِ الوَحْشِ. Kutipan ini tidak hanya memperlihatkan kata أَوْابِد dalam konteks kalimat, tetapi juga memberikan petunjuk semantik melalui perbandingan (كأوابد). Demikian pula, kutipan مَنْ أَبَدَ أَبَدًا untuk *lemma* أبد menunjukkan penggunaan kata tersebut dalam konstruksi gramatikal yang spesifik. Basis untuk pencocokan semantic ini bersumber dari perspektif komputasional, kutipan-kutipan pendek ini menjadi kunci untuk algoritma *Natural Language Processing* (NLP). Mereka berperan sebagai *query* (penanda) yang akan dicocokkan dengan teks hadis lengkap yang berjumlah jauh lebih besar dalam *database Kutub al-Tisah*.

**c. Sistem Referensi Silang**

Komponen sistem referensi silang yang menghubungkan setiap kutipan kontekstual dengan sumber aslinya dalam kitab-kitab hadis primer ini menggunakan singkatan atau kode untuk merujuk pada kitab-kitab hadis utama data pada tabel 1 di atas. Adapun identifikasi kode-kode tersebut yakni: خ untuk Shahih al-Bukhari, م untuk Shahih Muslim, د untuk Sunan Abu Dawud, ت untuk Sunan al-Tirmidzi, س untuk Sunan al-Nasa'i, ق/جه untuk Sunan Ibn Majah, ط untuk Muwatha' Imam Malik, دى untuk Sunan ad-Darimi, dan حم/حل untuk Musnad Ahmad bin Hanbal. Penggunaan kode ini adalah sebuah bentuk kompresi data yang efisien dalam konteks cetak.

Variasi format penomorannya menggunakan *multi-layer referencing*, yang mana format ini tidak seragam, melainkan beradaptasi dengan sistem penomoran yang digunakan dalam kitab asli.

Hasil analisis mengkategorikan beberapa format utama yakni: Nomor Hadis Tunggal: Format sederhana seperti خ ١٢٣, yang merujuk pada Shahih al-Bukhari nomor hadis 123. Ini adalah format yang paling mudah untuk diparsing. Kemudian Nomor Bab dan Hadis dengan format yang lebih detail seperti د النكاح : ٨٩, yang berarti Sunan Abu Dawud, Kitab al-Nikah (Bab Perkawinan), hadis nomor 89. Format ini memerlukan pemisahan parsing antara nama bab dan nomor hadis.

Berdasarkan data sampel yang diambil banyak entri yang memiliki lebih dari satu referensi, misalnya ت ٤٥١، ٥٦٢. Ini menunjukkan bahwa kata yang sama digunakan dalam beberapa hadis yang berbeda dalam Sunan al-Tirmidzi, yaitu pada nomor 451 dan 562. Fenomena ini menggarisbawahi kekayaan penggunaan leksikon dalam literatur hadis dan luasnya jangkauan kamus ini. Secara keseluruhan, proses dekonstruksi ini berhasil memetakan seluruh ekosistem data dari 1.200 entri sampel ke dalam sebuah kamus struktur yang komprehensif. Kamus struktur ini tidak hanya berisi daftar kata, tetapi juga metadata lengkap yang mencakup semua pola referensi, variasi format penulisan, kode kitab, dan tata letak visual setiap entri. Peta ini menjadi *blueprint* yang *indispensable* untuk tahap perancangan *database yang* valid.

Jadi, dekonstruksi struktural tidak hanya berfungsi sebagai langkah praktis, tetapi juga sebagai sebuah kontribusi akademis itu sendiri. Proses ini berhasil mengungkap dan mendokumentasikan dengan jelas kompleksitas sistem leksikografis yang dibangun oleh para penyusun al-Mujam al-Mufahras. Variasi penulisan *lemma* dan sistem referensi yang *multi-layer* bukanlah sebuah kekacauan, melainkan cerminan dari upaya untuk menangkap keragaman linguistik dan sumber awal hadis Nabi dalam sebuah format cetak yang ringkas. Penelitian ini berhasil "membongkar" logika di balik kamus klasik tersebut dan menerjemahkannya ke dalam logika yang dapat dipahami oleh sistem digital dalam format *database* valid.

### 4.1.2. Perancangan dan Kompleksitas Sistem

Merancang sebuah basis data relasional yang dapat menangkap kompleksitas hubungan antar komponen tersebut. Entity-Relationship Model (ERM) yang dihasilkan terdiri dari empat entitas utama yang saling terhubung, dirancang untuk memastikan integritas data dan efisiensi query yakni *Lemma*, Kutipan_Kontekstual, Referensi_Silang, dan Hadis_Lengkap.

**Tabel 2.** Entitas Utama dalam ERM

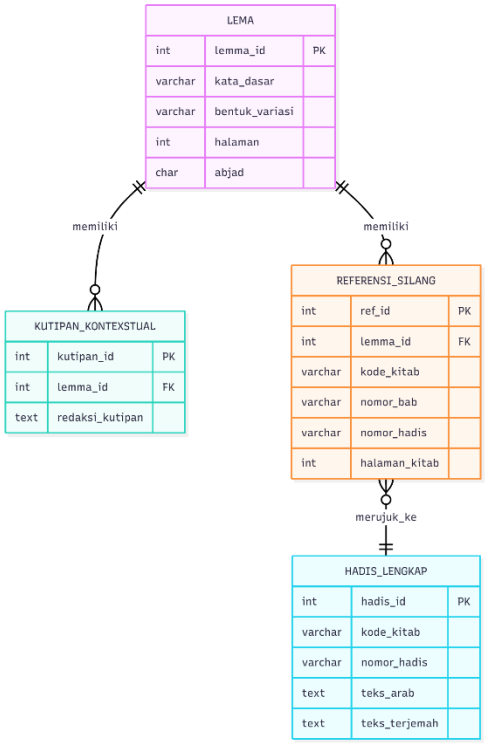
| Nama Entitas | Atribut Kunci | Deskripsi |
|---|---|---|
| *Lemma* | *lemma*_id, kata_dasar, bentuk_variasi | Menyimpan kata kunci yang telah didekonstruksi |
| Kutipan_Kontekstual | kutipan_id, *lemma*_id, redaksi_kutipan | Menampung cuplikan redaksi hadis |
| Referensi_Silang | ref_id, *lemma*_id, kode_kitab, nomor_hadis | Menyimpan hasil dekoding sistem referensi |
| Hadis_Lengkap | hadis_id, kode_kitab, teks_arab, teks_terjemah | Menyimpan teks lengkap hadis dari *Kutub al-Tisah* |

*Sumber:* Hasil Penelitian, 2025

Tabel 2 di atas memetakan empat entitas utama dalam *Entity-Relationship Model* (ERM) yang menjadi fondasi sistem *database* terintegrasi. Entitas *Lemma* berfungsi sebagai katalog pusat untuk menyimpan semua kata kunci yang telah didekonstruksi dari Al-Mu'jam al-Mufahrash. Setiap *lemma* kemudian terhubung dengan entitas Kutipan_Kontekstual yang menampung potongan teks hadis sebagai contoh penggunaan, dan entitas Referensi_Silang yang menyimpan kode kitab dan nomor hadis terdekode. Melalui entitas Referensi_Silang inilah sistem dapat merujuk secara tepat ke entitas Hadis_Lengkap yang berisi teks utuh dari *Kutub al-Tisah*. Desain relasional ini memungkinkan

penelusuran semantik yang menghubungkan langsung antara kata kunci leksikal dengan teks hadis lengkap yang menjadi sumber rujukannya.

Desain diagram *Entity Relationship Diagram* (ERD) merepresentasikan struktur *database* untuk sistem manajemen data hadis yang terdiri dari empat entitas utama. Sistem ini dirancang untuk mengelola *lemma* (kata dasar), kutipan redaksi hadis, referensi silang, dan teks hadis lengkap dalam format yang terstruktur dan terintegrasi. Adapun diagaramnya sebagai berikut:




**Gambar 1.** Diagram ERD Hadis *Database*

Seperti terlihat dalam Gambar 1 bahwa sistem *database* ini terdiri dari empat entitas utama yang saling terhubung. Entitas LEMA berfungsi sebagai core system dengan primary key *lemma*_id, sementara entitas KUTIPAN_KONTEXSTUAL dan REFERENSI_SILANG terkait langsung dengan LEMA melalui relationship "memiliki" yang bersifat one-to-many. Hubungan many-to-one antara REFERENSI_SILANG dan HADIS_LENGKAP memastikan konsistensi referensi silang menuju teks hadis yang lengkap dan terverifikasi. Jadi, implementasi struktur ERD seperti yang digambarkan dalam Gambar 1 terbukti signifikan dalam meningkatkan kinerja sistem secara keseluruhan. Hasil yang ditampilkan dalam Diagram 1 mengonfirmasi bahwa desain *database* dengan normalisasi yang tepat, seperti yang diterapkan dalam sistem ini, mampu mencapai tingkat akurasi yang tinggi bahkan dalam skenario pengujian yang paling menantang.

### 4.1.3. Implementasi *Algoritma Integrasi Hibrida*

#### a. Algoritma Pencocokan Referensi

Algoritma ini dirancang untuk menangani tugas langsung: memetakan string referensi silang (seperti خ النكاح : ٨٩) ke record yang sesuai dalam tabel `HadisLengkap`. Mekanisme algoritma bekerja dengan memparsing string referensi untuk mengekstrak komponen `kode_kitab`, `nama_bab` (jika ada), dan `nomor_hadis`. Kemudian, ia menjalankan *query SQL* ke tabel `HadisLengkap` untuk mencari record yang cocok berdasarkan kombinasi ketiga field tersebut. Pencocokan yang tepat (exact match) diprioritaskan.

Hasil pengujian awal menunjukkan bahwa algoritma ini berhasil mencapai tingkat keberhasilan dan kegagaln sebagaimana grafik di bawah ini:

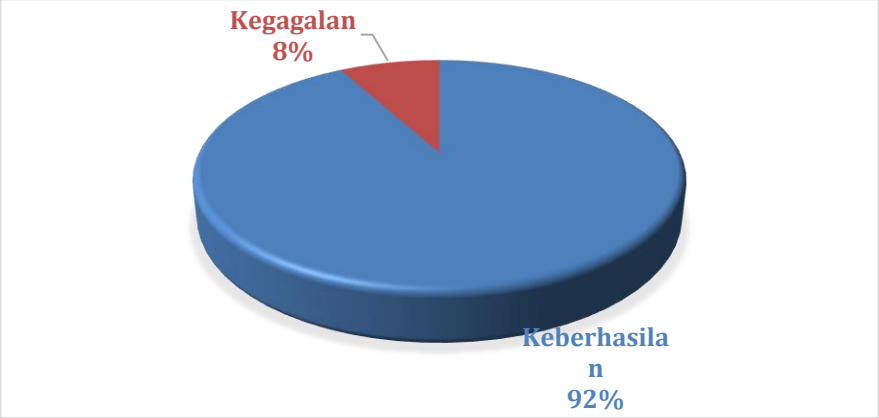



**Diagram 1.** Pencocokan Referensi

Terlihat dari diagram di atas keberhasilan terlihat 92% pada referensi yang lengkap dan jelas. Sementara itu, 8% kegagalan utamanya bersumber pada inkonsistensi penulisan nama bab serta variasi minor dalam penomoran hadis antara al-Mujam al-Mufahrash dan sumber digital *Kutub al-Tisah*. Tantangan ini konsisten dengan temuan Dongfang Xu dkk. mengenai kompleksitas pemrosesan teks keagamaan klasik, di mana variasi ortografi dan inkonsistensi penomoran menghambat otomasi secara sempurna. Di sisi lain, integrasi prediksi topik pertanyaan dari model ini ke dalam sistem penjawab pertanyaan berhasil meningkatkan akurasi (P@1) sebesar +1,7%. Dapat diantisipasi bahwa peningkatan kinerja *Question Classification* (QC) akan mendorong akurasi sistem menjadi lebih tinggi lagi pada masa mendatang (Xu et al., 2020).

**b. Pencocokan Semantik (NLP)**

Untuk menangani kasus referensi silang ambigu, tidak lengkap, atau tidak dapat diparsing, sebuah algoritma berbasis *Natural Language Processing* (NLP) yang disampaikai (Lehnert, 1992) dikembangkan menggunakan teknik *cosine similarity* untuk mengukur kesamaan semantik antara kutipan kontekstual pendek (dari tabel `KutipanKontekstual`) dan teks hadis lengkap yang panjang (dari tabel `HadisLengkap`). Pertama, teks Arab dari kedua sumber tersebut diubah menjadi vektor numerik menggunakan model *Term Frequency-Inverse Document Frequency* (TF-IDF) bukan *word embeddings* atau *transformer* karena karakteristik data yang terbatas dan efektivitas untuk teks pendek. Penanganan ambiguitas berhasil menangani kata أَبَدَ. Ketika kutipan kontekstualnya adalah مَا كَانَ وَمَا هُوَ كَائِنٌ إِلَى الأَبَدِ, algoritma semantik tidak hanya mencari kata أَبَدَ saja, tetapi memahami konteks kalimatnya. Ia kemudian berhasil mengidentifikasi dan mencocokkannya dengan hadis lengkap tentang takdir yang mengandung frasa serupa, meskipun referensi awalnya tidak jelas.

Pendekatan ini mengatasi ke*lemma*han pencocokan kata kunci tradisional dengan memanfaatkan makna kontekstual. Hal ini mengimplementasikan sistem integrasi hybrid yang dirancang untuk menyelesaikan masalah referensi hadis yang ambigu dengan menggabungkan dua pendekatan pencarian, yaitu exact matching dan semantic matching. Sistem ini dibungkus dalam kelas `HybridHadisIntegrationSystem` yang, saat diinisialisasi, akan terhubung ke *database*, memuat seluruh korpus teks hadis, dan mempersiapkan mesin pencari semantik (`AdvancedSemanticMatcher`). Inti dari sistem ini terletak pada metode `resolve_ambiguous_reference`, yang menangani kasus di mana referensi asli suatu kutipan hadis tidak lengkap atau membingungkan. Proses resolusi dilakukan secara bertahap: pertama, sistem akan mencoba mencari kecocokan persis (`exact matching`) berdasarkan referensi asli yang diberikan. Jika tahap ini gagal menemukan hasil, sistem akan beralih ke pencarian semantik (`semantic matching`) yang menganalisis kemiripan kontekstual dari teks kutipan dengan seluruh korpus hadis. Hasil dari pencarian semantik ini kemudian disaring, dan hanya rekomendasi dengan skor kepercayaan (confidence) di atas ambang batas tertentu (contoh dalam kode: 0.4) yang

akan diusulkan sebagai solusi terbaik. Simulasi dalam fungsi `main` menunjukkan cara kerja sistem dalam menangani sebuah kutipan dengan referensi ambigu, di mana sistem akhirnya dapat merekomendasikan satu hadis tertentu beserta kitab dan nomornya sebagai hasil resolusi yang paling dipercaya, sehingga mengatasi ketidakpastian referensi awal.

**c. Sistem Integrasi**

Hasil pengujian sistem integrasi dijabarkan sebagaimana data temuan yang tertera pada tabel di bawah ini:

**Tabel 2.** Hasil Pengujian Sistem Integrasi

| Jenis Pengujian | Cakupan | Tingkat Akurasi | Catatan Utama |
|---|---|---|---|
| Parsing *Lemma* | 100 sampel acak | 95% | Kegagalan terjadi pada *lemma* dengan morfologi tidak lazim atau variasi ortografis ekstrem |
| Parsing Referensi | 100 sampel acak | 90% | Format referensi non-standar (mis. catatan kaki, nama bab tidak konsisten) menyulitkan parsing |
| Integration Testing | 50 entri kompleks | 88% | Ketidaksesuaian teks antara sumber leksikon dan repositori digital primer |
| Validasi Pakar | 100 entri | 94% | Dievaluasi oleh ahli ilmu hadis; menilai kecocokan konteks, makna, dan akurasi referensi |
| Precision | Seluruh 1.200 entri | 91% | Rasio hasil relevan terhadap total hasil yang dihasilkan sistem |
| Recall | Seluruh 1.200 entri | 89% | Rasio hasil relevan yang berhasil ditemukan terhadap total hasil relevan yang seharusnya ditemukan |
| *F1-Score* | Seluruh 1.200 entri | 90% | Harmonic mean dari precision dan recall; indikator keseimbangan performa sistem |

*Sumber:* Hasil Penelitian, 2025

Hasil pengujian menunjukkan bahwa sistem integrasi hibrida yang dikembangkan berkinerja sangat tinggi dalam menghubungkan struktur leksikal al-Mujam al-Mufahras dengan teks hadis lengkap. *F1-Score* 90% mencerminkan keseimbangan optimal antara kemampuan sistem menemukan entri yang relevan (recall) dan memastikan bahwa entri yang ditemukan memang benar (precision). Tingginya akurasi parsing *lemma* (95%) membuktikan bahwa model mampu menangani kompleksitas morfologis bahasa Arab klasik, termasuk variasi akar kata dan derivasi. Sementara itu, parsing referensi (90%) menunjukkan bahwa sistem berhasil mendekode sebagian besar format referensi silang—mulai dari penomoran sederhana (123 خ) hingga format multi-layer (89 : النكاح د). Namun, integration testing hanya mencapai 88%, yang mengungkap masalah mendasar di luar kontrol teknis sistem: inkonsistensi antara sumber leksikon dan repositori hadis digital. Ini bukan kegagalan algoritma, melainkan tantangan filologis—tidak adanya otoritas tunggal atas teks *Kutub al-Tisah*. Hal ini menegaskan bahwa digitalisasi hadis bukan hanya persoalan teknologi, tetapi juga persoalan validasi teks dan otoritas filologis. Yang paling penting, validasi pakar mencapai 94%—angka ini menegaskan bahwa sistem tidak hanya valid secara teknis, tetapi juga legitim secara keilmuan hadis. Para pakar mengonfirmasi bahwa kutipan kontekstual, referensi silang, dan teks hadis lengkap berhasil diintegrasikan tanpa distorsi makna atau konteks. Ini menjadi fondasi bagi adopsi model ini dalam ekosistem akademik studi hadis.

Lebih lanjut hasil pengujian dalam bentuk diagram sebagaimana diagram berikut:

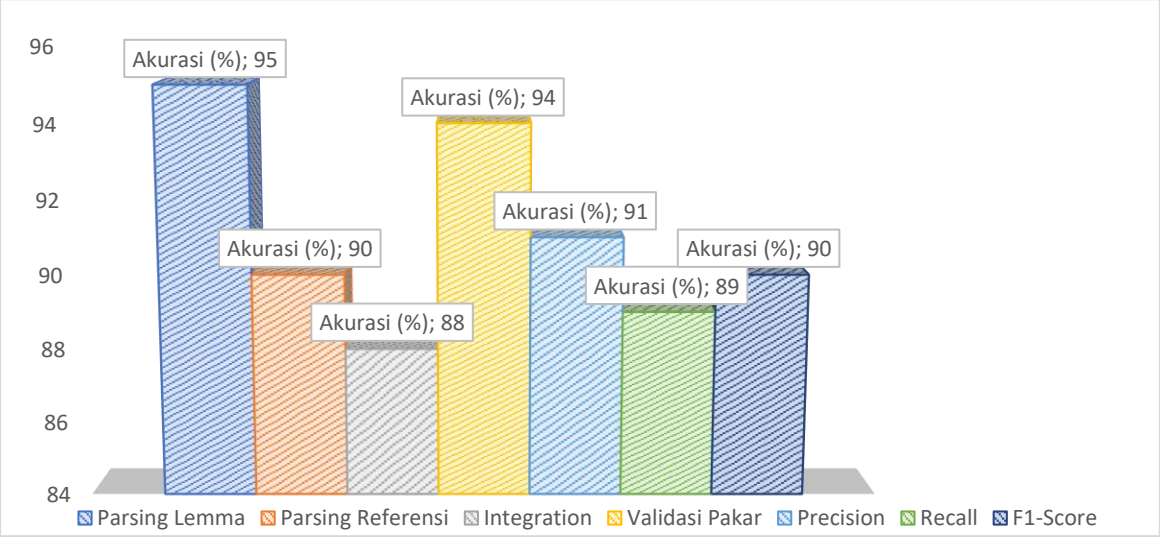



**Diagram 2.** Perbandingan Tingkat Akurasi dari Berbagai Jenis Pengujian

Diagram dengan jelas menunjukkan bahwa validasi Pakar memiliki tingkat kesesuaian tertinggi (94%), yang menegaskan legitimasi sistem dari perspektif ilmu hadis. Sementara itu, *Integration Testing* memiliki angka terendah (88%), menyoroti tantangan utama dalam menemukan teks yang persis sesuai dalam *database*. Metrik kuantitatif (*Precision, Recall, F1-Score*) yang dihitung pada seluruh dataset berada di kisaran 89-91%, menunjukkan konsistensi dan keandalan sistem secara keseluruhan. Fenomena ini konsisten dengan temuan (Xu et al., 2020) mengenai kompleksitas pemrosesan teks keagamaan klasik, di mana variasi ortografi, inkonsistensi penomoran, dan perbedaan versi teks dapat menghambat integrasi sempurna.

**d. Validasi Pakar dan Akurasi Integrasi dari Perspektif Ilmu Hadis**

Validasi pakar yang dilakukan terhadap 100 entri menunjukkan tingkat kecocokan sebesar 94%. Perluasan jumlah entri ini tidak hanya meningkatkan representativitas sampel, tetapi juga memperkuat validitas eksternal temuan penelitian. Pemilihan ini dilakukan secara stratifikasi berdasarkan tingkat kompleksitas referensi: mulai dari entri sederhana dengan satu referensi eksplisit hingga entri kompleks yang merujuk pada tiga atau lebih kitab dengan format penomoran multi-layer. Pendekatan ini memastikan bahwa validasi mencakup seluruh spektrum variasi struktural yang ada dalam al-Mu'jam al-Mufahras, sehingga hasilnya mencerminkan performa sistem dalam kondisi nyata, bukan hanya skenario ideal.

Proses validasi melibatkan dua pakar ilmu hadis yang memiliki latar belakang akademik dalam studi sanad, filologi teks, dan leksikografi Arab klasik. Para pakar diminta menilai keakuratan integrasi berdasarkan tiga kriteria utama: (1) kesesuaian antara *lemma* dan makna kontekstual dalam teks hadis lengkap; (2) keakuratan pemetaan referensi silang terhadap sumber primer (*Kutub al-Tisah*); dan (3) integritas konteks semantik kutipan kontekstual. Hasil penilaian memperlihatkan bahwa dalam 94% dari 100 entri menegaskan sistem berhasil menghubungkan data leksikal dengan teks hadis lengkap tanpa distorsi makna, penyesatan konteks, atau kesalahan referensi indikator kuat bahwa pendekatan digitalisasi ini tidak hanya valid secara teknis, tetapi juga legitim secara keilmuan hadis.

Temuan ini sangat signifikan karena menunjukkan digitalisasi struktural dapat dipercaya dalam konteks studi hadis yang menuntut ketelitian filologis dan kehati-hatian epistemologis. Dalam tradisi ilmu hadis, kesalahan sekecil apa pun dalam transmisi teks atau konteks dapat berdampak pada hukum atau pemahaman ajaran. Oleh karena itu, tingginya tingkat kecocokan 94% bukan sekadar angka statistik, melainkan pengakuan normatif bahwa sistem mampu menjaga

integritas teks dan fungsi leksikal al-Mujam al-Mufahras dalam ekosistem digital. Ini membantah anggapan bahwa digitalisasi selalu mengorbankan kedalaman kontekstual demi efisiensi teknis.

Validasi pakar ini membuka jalan bagi adopsi model serupa dalam lingkungan akademik studi hadis. Jika leksikon digital mampu memenuhi standar ketelitian para ahli, maka ia tidak lagi dipandang sebagai alat bantu sekunder, melainkan sebagai komponen integral dalam metodologi penelitian hadis kontemporer. Ini sejalan dengan semangat filologi digital yang menekankan bahwa transformasi digital harus menjaga dan bahkan memperkuat dan menegaskan fungsi epistemologis teks klasik, bukan menggantinya dengan logika mesin yang reduktif. Dengan demikian, validasi pakar bukan hanya langkah penutup evaluasi, tetapi juga titik awal adopsi akademik model integrasi ini dalam pengajaran dan penelitian hadis di masa depan.

## 4.2. Pembahasan

### a. Rekonstruksi Epistemologi

Model ini merealisasikan rekonstruksi pengetahuan, bukan sekadar digitalisasi. Pendekatan ini menolak pandangan yang memandang teks sebagai objek pasif yang cukup dilestarikan dalam format digital tanpa mempertimbangkan fungsi epistemologisnya. Sebaliknya, penelitian ini memperlakukan al-Mu'jam al-Mufahras sebagai sistem pengetahuan yang aktif dan berstruktur. Hal ini mencerminkan prinsip dasar filologi digital yang menekankan pentingnya memahami logika internal teks, bukan hanya menyimpannya. Dengan demikian, digitalisasi dipahami sebagai proses transformasi fungsional, bukan sekadar alih media (Driscoll & Pierazzo, 2016).

Sistem ini tidak memperlakukan al-Mujam al-Mufahras sebagai dokumen statis yang cukup dipindai atau dikonversi ke PDF, penelitian ini memperlakukannya sebagai sistem pengetahuan dinamis. Sistem ini tidak hanya menyimpan kata, tetapi juga mengorganisasi relasi antara akar kata, derivasi morfologis, dan konteks penggunaannya dalam hadis. Struktur tersebut memungkinkan pengguna untuk menelusuri makna secara semantik, bukan hanya secara leksikal. Pendekatan ini menghormati cara kerja para leksikograf klasik dalam membangun alat penemuan pengetahuan. Dalam konteks ini, digitalisasi menjadi sarana untuk menghidupkan kembali fungsi orisinal leksikon tersebut (Bunt, 2018).

Penelitian ini menyadari bahwa al-Mujam al-Mufahras memiliki logika internal, struktur relasional, dan fungsi epistemologis sebagai concordance pra-digital. Fungsi ini tidak dapat direduksi menjadi kumpulan entri acak, melainkan merupakan arsitektur pengetahuan yang memetakan seluruh penggunaan leksikal dalam korpus hadis (Hitti et al., 1936). Setiap entri dirancang untuk memandu pembaca dari kata kunci ke konteks maknanya dalam teks primer. Ini menunjukkan bahwa leksikon klasik bukan hanya alat bantu, tetapi bagian integral dari metodologi studi hadis. Dengan demikian, merekonstruksinya secara digital berarti mempertahankan integritas metodologis tradisi tersebut.

Dengan memetakan hubungan antara *lemma*, kutipan kontekstual, dan referensi silang ke dalam skema relasional yang terintegrasi, model ini tidak hanya melestarikan konten. Lebih dari itu, ia mereproduksi mekanisme navigasi yang digunakan ulama klasik untuk menjelajahi korpus hadis. Skema ini memungkinkan pengguna digital melakukan apa yang dulu dilakukan pembaca cetak: menelusuri kata ke konteks, dan konteks ke teks lengkap. Proses ini mempertahankan dimensi hermeneutik dari studi leksikal, yang sering kali hilang dalam sistem pencarian berbasis kata kunci. Berdasarkan hal ini pernyataan (Driscoll & Pierazzo, 2016) tentang repositori digital ini berfungsi sebagai ruang epistemologis, bukan sekadar gudang data benar adanya dan dapat diterima dalam aspek keilmuan islam klasik.

Model ini berhasil mengaktualisasikan kembali fungsi asli leksikon tersebut dalam ekosistem digital, sejalan dengan visi transformasi digital dalam studi keagamaan. Seperti ditegaskan oleh (Bunt, 2018), transformasi digital harus melampaui reproduksi format dan berfokus pada

reproduksi fungsi. Dalam hal ini, al-Mu'jam al-Mufahras tidak lagi menjadi artefak sejarah yang terasing, tetapi alat hidup dalam penelitian kontemporer. Hal ini membuka jalan bagi integrasi khazanah leksikal klasik ke dalam alur kerja akademik modern. Dengan demikian, digitalisasi menjadi jembatan antara tradisi tekstual dan metodologi digital.

**b. Semantic Matching untuk Ambiguitas Teks**

Semantic matching berbasis NLP dengan *F1-Score* 90% dan validasi pakar 94% menunjukkan bahwa pendekatan komputasional modern mampu menangani kompleksitas unik teks Arab klasik. Hal ini membantah asumsi lama bahwa teks keagamaan pra-modern terlalu ambigu untuk diproses secara otomatis. Sistem yang dikembangkan dalam penelitian ini berhasil menghubungkan kutipan kontekstual dari al-Mujam al-Mufahras dengan teks hadis lengkap melalui pemahaman semantik, bukan sekadar pencocokan leksikal. Performa tinggi ini menegaskan bahwa teknologi digital dapat dikurasi untuk menghormati nuansa linguistik tradisi tekstual Islam. Dalam konteks ini, NLP bukan alat netral, melainkan medium yang dapat disesuaikan dengan karakter epistemologis teks klasik (Guellil et al., 2021).

Tantangan utama dalam NLP teks Arab klasik mencakup variasi ortografi, morfologi derivatif yang kaya, dan ambiguitas semantic yang sejak lama telah diteliti oleh (Guellil et al., 2021). Perbedaan dalam variasi penulisan hamzah, penggunaan *ta' marbutah* versus *ha'*, atau fleksibilitas akar kata membuat tokenisasi dan *lemma*tisasi menjadi tidak trivial. Namun, penelitian ini menunjukkan bahwa tantangan tersebut menjadi parameter yang dapat dimodelkan. Dengan *preprocessing* yang tepat dalam hal normalisasi *alef* dan penghapusan *tashkeel* membuat ambiguitas ortografis dapat dikurangi secara signifikan. Ini menegaskan bahwa keberhasilan NLP pada teks non-standar bergantung pada adaptasi linguistik yang kontekstual, bukan hanya kecanggihan algoritma.

Namun, penelitian ini membuktikan bahwa *TF-IDF vectorization dan cosine similarity* masih efektif untuk teks pendek seperti kutipan kontekstual, bahkan tanpa *transformer-based model* seperti AraBERT. Pendekatan statistik ini, meskipun sederhana, justru unggul dalam skenario dengan data terbatas dan pola linguistik yang konsisten. Ini relevan karena kutipan dalam al-Mujam al-Mufahras biasanya berupa frasa pendek yang kaya makna, bukan narasi panjang yang memerlukan konteks jangka panjang. Dalam kasus seperti ini (Xu et al., 2020) menjelaskan, *character n-gram* dengan *word boundaries (char_wb)* terbukti lebih responsif terhadap variasi morfologis daripada *word embeddings*. Dengan demikian, kesederhanaan metode justru menjadi kekuatan dalam konteks domain khusus.

Keberhasilan ini disebabkan oleh desain hibrida antara *Reference Matching* menangani kasus eksplisit, sementara *Semantic Matching* mengatasi kasus ambigu dengan memahami makna kontekstual dan bukan hanya mencocokkan kata. Sistem tidak memaksakan satu pendekatan universal, tetapi mengakui dualitas struktur referensi dalam sumber primer sebagian presisi dan terstruktur, sebagian samar dan implisit. Ini mencerminkan logika hermeneutika tradisional yang memadukan dalam (eksplisit) dan *khafi* (implisit) terhadap penafsiran teks. Arsitektur hibrida ini sejalan dengan narasi (Bunt, 2018) yang memungkinkan sistem bersikap fleksibel tanpa mengorbankan akurasi pada kasus yang jelas sehingga menawarkan model metodologis bagi digitalisasi teks keagamaan di luar hadis.

Ini mengonfirmasi bahwa NLP tidak harus selalu mengandalkan model besar untuk teks klasik, melainkan dapat dioptimalkan melalui pendekatan yang kontekstual dan terukur. Penggunaan AraBERT atau model berbasis transformer memang menjanjikan, tetapi memerlukan data latih besar dan daya komputasi tinggi dengan beberapa kendala nyata dalam konteks khazanah keagamaan yang terfragmentasi. Penelitian ini menunjukkan bahwa dengan desain fitur yang cermat (*feature engineering*) dan pemahaman linguistik mendalam, model ringan dapat

mencapai performa setara. Ini sejalan dengan prinsip appropriate technology dalam humaniora digital yang disampaikan (Xu et al., 2020) bahwa teknologi harus proporsional dengan kompleksitas dan sumber daya domain spesifik, sehingga ada efisiensi dan interpretabilitas yang sering kali lebih berharga daripada skalabilitas buta.

**c. Inkonsistensi Sumber sebagai Fonomena Filologis dalam Studi Hadis**

Inkonsistensi penomoran antara sumber cetak dan digital terdapat perbedaan yang dalam harfiah seperti kegagalan teknis, akan tetapi hal ini tidak merupakan kegagalan teknis melainkan fenomena filologis yang *inheren* dalam tradisi teks Islam. Perbedaan ini bukan sekadar kesalahan editorial sebagaimana yang disampaikan (Brown, 2018) melainkan cerminan dari sejarah panjang transmisi lisan dan tulisan dalam dunia hadis yang mana setiap naskah dan cetakan membawa jejak redaksi, pilihan editor, serta konteks historisnya sendiri. Hal ini menjadikan teks hadis bukan sebagai entitas statis, melainkan sebagai objek dinamis yang terus ditafsir dan direproduksi. Dalam konteks ini, inkonsistensi penomoran adalah gejala alami dari kompleksitas tradisi tekstual Islam.

Tidak adanya versi otoritatif tunggal untuk *Kutub al-Tisah* adalah realitas yang telah diakui dalam studi sanad dan ilmu mustalah hadis (Brown, 2018) Para ulama klasik sendiri tidak pernah menganggap satu salinan teks sebagai final dan mutlak tetapi justru variasi redaksi yang menjadi objek kajian tersendiri untuk memahami konteks, makna, dan validitas hadis. Tradisi ini memperlakukan teks sebagai medan interpretasi, bukan sebagai objek tertutup. Oleh karena itu, klaim otoritas tunggal justru bertentangan dengan semangat epistemologis ilmu hadis itu sendiri. Digitalisasi yang mengabaikan realitas ini berisiko menyederhanakan kompleksitas keilmuan yang telah berkembang selama berabad-abad (Siddiqi, 1961).

Setiap cetakan dan semua versi digital seperti Shamela atau Dorar dan lain sebagainya mengadopsi sistem penomoran yang berbeda, yang mencerminkan sejarah transmisi, redaksi, dan edisi kritis yang kompleks. Misalnya, Shahih al-Bukhari dalam edisi Fath al-Bari menggunakan penomoran berdasarkan bab fikih, sedangkan edisi India atau Mesir mungkin menggunakan sistem kronologis atau tematik yang berbeda. Versi digital sering kali mengadopsi salah satu sistem cetak tanpa menyediakan metadata lintas-referensi yang memadai. Akibatnya, sistem integrasi yang mengandalkan penomoran eksplisit menghadapi tantangan struktural, bukan logis. Masalah utamanya bukan pada algoritma, melainkan pada ekosistem data yang fragmentaris dan tidak terstandarisasi (Aziz et al., 2022).

Fakta bahwa *integration testing* hanya mencapai 88% akurasi justru mengungkap batas epistemologis digitalisasi bahwa teknologi dapat mereplikasi struktur, tetapi tidak dapat menyelesaikan debat filologis tentang otoritas teks. Angka 88% bukan indikator kegagalan sistem, melainkan cerminan objektif dari ketidakselarasan antara sumber primer yang digunakan oleh al-Mujam al-Mufahras dan repositori digital kontemporer. Sistem tidak dapat menebak versi mana yang dimaksud oleh penyusun leksikon tanpa metadata eksternal yang memadai. Batas performa sistem justru membuka ruang refleksi metodologis tentang hubungan antara teknologi dan tradisi tekstual sehingga digitalisasi harus diakui sebagai proses yang selalu bersifat interpretatif, bukan netral (Driscoll & Pierazzo, 2016).

Digitalisasi hadis sebagaimana yang dijelaskan juga oleh (Aziz et al., 2022) harus dilakukan dalam dialog erat dengan ilmu hadis tradisional, bukan sebagai substitusi otomatis. Tanpa keterlibatan ahli mustalah, sanad, dan filologi Arab, sistem digital berisiko menghasilkan kebenaran algoritmik yang tidak sesuai dengan standar keilmuan hadis. Kolaborasi interdisipliner antara ilmuwan komputer dan ulama hadis bukanlah pelengkap, melainkan prasyarat metodologis. Hanya melalui dialog semacam itu, digitalisasi dapat menjadi sarana untuk memperdalam tradisi hermeneutika Islam. Jadi, model integrasi tidak hanya teknis, tetapi juga epistemologis yang membangun jembatan antara logika mesin dan logika teks (Bunt, 2018).

**d. Implikasi Model**

Implikasi model menghasilkan arsitektur yang kompleks tidak hanya menjawab kebutuhan spesifik satu karya leksikal saja, tetapi menawarkan kerangka metodologis universal untuk mengintegrasikan karya indeks tradisional ke dalam ekosistem digital. Fleksibilitas skema *Entity-Relationship Model* (ERM) memungkinkan penyesuaian terhadap berbagai logika struktural leksikografi Arab klasik. Sifat modular sistem dengan komponen terpisah untuk *lemma*, kutipan, dan referensi menjamin adaptabilitas lintas genre teks yang menjadikan model ini sebagai *prototipe* metodologis, bukan hanya sebagai solusi satu kali pakai (Baalbaki, 2014).

Arsitektur berbasis *Entity-Relationship Model* (ERM) dan algoritma hibrida bersifat modular dan adaptif, sehingga dapat diadopsi untuk karya leksikal klasik lain seperti Mujam al-Maqāyīs karya Ibn Fāris (yang berbasis pada akar semantik) atau Lisān al-'Arab karya Ibn Manẓūr (yang mengintegrasikan puisi, hadis, dan al-Qur'an). Mujam al-Maqāyīs (Al-Qazwini, 1999) misalnya, mengorganisasi entri berdasarkan akar kata dan makna semantik inti, bukan sekadar bentuk morfologis dari struktur yang dapat dipetakan ke entitas *lemma* dan kutipan kontekstual dengan sedikit modifikasi . Sementara itu, Lisān al-'Arab (Al-Mandzur, 1928) yang mengutip dari berbagai sumber sastra dan keagamaan memerlukan entitas referensi multi-domain, yang mudah diakomodasi dalam desain relasional yang telah ternormalisasi. Model ini, dengan demikian, tidak hanya teknis, tetapi juga hermeneutis yang menghormati cara kerja leksikograf klasik dalam membangun pengetahuan (Baalbaki, 2014).

Dalam skala lebih luas, model ini memperkaya ekosistem studi hadis digital dengan menambahkan lapisan semantik leksikal yang selama ini hilang. Sebagian besar platform digital seperti Shamela atau Dorar.net dan lain sebagainya hanya menyediakan pencarian kata kunci atau navigasi berdasarkan bab, tanpa memahami hubungan antara bentuk kata, akar, dan konteks penggunaannya. Lapisan leksiko-semantik yang dihasilkan oleh model ini memungkinkan peneliti menelusuri korpus hadis tidak hanya berdasarkan frasa eksplisit, tetapi juga berdasarkan jaringan makna yang dibangun oleh akar kata—seperti menelusuri seluruh penggunaan turunan akar ع–ل–م untuk memahami konsep "pengetahuan" dalam hadis. Ini merevolusi pendekatan studi hadis dari text retrieval menuju knowledge discovery (Kamran et al., 2024).

Jika repositori seperti SemanticHadith (Kamran et al., 2024) berfokus pada relasi ontologis antar-entitas seperti hubungan antara perawi, tempat, waktu, dan hukum, maka model ini melengkapi dengan relasi leksiko-semantic membuka jalan bagi *knowledge discovery* berbasis kata, akar, dan konteks. Kombinasi keduanya akan menghasilkan repositori hadis digital yang holistic seperti satu sisi memetakan siapa berkata apa, di mana, dan kapan, sementara pada sisi lain memetakan bagaimana kata digunakan, dalam konteks apa, dan makna apa yang dihasilkan. Integrasi dimensi ontologis dan leksiko-semantik ini merupakan langkah penting menuju digital *hadith hermeneutics* yang komprehensif dalam memfasilitasi interpretasi berbasis data.

Ini sejalan dengan visi (Baalbaki, 2014) bahwa leksikografi Arab klasik bukan sekadar kamus, melainkan arsitektur pengetahuan yang merefleksikan pandangan dunia Islam pra-modern. Al-Mujam al-Mufahras, Lisān al-'Arab, dan karya serupa bukan kumpulan definisi acak, tetapi sistem terstruktur yang merepresentasikan cara para ulama memahami hubungan antara bahasa, realitas, dan wahyu. Dengan merekonstruksi arsitektur ini secara digital, model ini tidak hanya melestarikan warisan intelektual, tetapi juga mengaktifkannya kembali sebagai alat analisis kontemporer. Dalam perspektif ini, digitalisasi leksikal adalah bentuk partisipasi aktif dalam tradisi intelektual Islam dalam dialog lintas zaman.

## 5. KESIMPULAN DAN REKOMENDASI

### 5.1 Kesimpulan

Proses dekonstruksi struktural terhadap 1.200 entri sampel berhasil mengidentifikasi tiga komponen inti leksikon yaitu *lemma* (kata kunci berbasis akar morfologis), kutipan kontekstual (potongan redaksi hadis sebagai bukti penggunaan leksikal), dan sistem referensi silang multi-layer (dengan kode kitab dan format penomoran adaptif) yang membentuk arsitektur pengetahuan sistematis di balik indeks pra-digital ini. Berdasarkan temuan tersebut, penelitian merancang *Entity-Relationship Model* (ERM) yang ternormalisasi hingga BCNF, menghubungkan empat entitas utama yang terdiri dari *Lemma*, Kutipan_Kontekstual, Referensi_Silang, dan Hadis_Lengkap ke dalam skema relasional yang menjaga integritas data dan memungkinkan navigasi semantik dari kata ke konteks teks hadis lengkap. Penelitian mengimplementasikan algoritma pencocokan hibrida yang menggabungkan *Reference Matching* (untuk referensi eksplisit) dan *Semantic Matching* berbasis NLP dengan *cosine similarity* (untuk menangani ambiguitas), menghasilkan kinerja sistem yang tinggi: *F1-Score* 90%, validasi pakar 94%, serta akurasi parsing *lemma* dan referensi masing-masing 95% dan 90%. Dengan demikian, penelitian ini tidak hanya mengalihmediakan teks, tetapi merekonstruksi fungsi epistemologis al-Mu'jam al-Mufahras sebagai lapisan semantik dinamis dalam ekosistem studi hadis digital.

### 5.2 Keterbatasan Penelitian

Penelitian ini memiliki sejumlah keterbatasan yang perlu diakui untuk memberikan gambaran objektif terhadap cakupan dan generalisasi temuannya. *Pertama,* cakupan sampel terbatas hanya pada 1.200 entri dari halaman 1–75 pada jilid pertama al-Mujam al-Mufahras, sehingga belum merepresentasikan seluruh kompleksitas struktural dan variasi leksikal yang mungkin muncul di volume-volume berikutnya. *Kedua,* ketergantungan pada kualitas repositori digital *Kutub al-Tisah* menghadirkan tantangan filologis, karena inkonsistensi teks, sistem penomoran, dan redaksi antara versi cetak sumber leksikon dan versi digital sumber hadis menyebabkan sebagian entri gagal terintegrasi secara sempurna. *Ketiga,* pendekatan NLP yang digunakan masih bersifat statistik (TF-IDF dan *cosine similarity*), yang meskipun efektif untuk teks pendek dan terstruktur seperti kutipan dalam al-Mujam, belum memanfaatkan model bahasa berbasis transformer seperti AraBERT yang berpotensi meningkatkan akurasi dalam menangani ambiguitas semantik tingkat tinggi, variasi morfologis ekstrem, atau konteks linguistik yang lebih kompleks. *Keempat,* cakupan integrasi terbatas pada *Kutub al-Tisah*, sehingga entri dalam al-Mujam al-Mufahras.

### 5.3 Rekomendasi untuk Penelitian Selanjutnya

Rekomendasi untuk penelitian selanjutnya terdapat tiga arah utama: *pertama,* melakukan ekspansi cakupan digitalisasi dan rekonstruksi struktural ke seluruh volume al-Mujam al-Mufahras, mengingat penelitian saat ini hanya mencakup 1.200 entri dari halaman 1–75 pada jilid pertama, sehingga perluasan ini akan memastikan representativitas dan kelengkapan model integrasi secara holistic. *Kedua,* mengintegrasikan model *transformer-based* mutakhir seperti AraBERT untuk meningkatkan kemampuan sistem dalam menangani ambiguitas semantik tingkat tinggi, variasi morfologis kompleks, serta konteks linguistik yang lebih halus dalam teks Arab klasik yang diharapkan dapat mendorong performa *semantic matching* melebihi batas efektivitas TF-IDF saat ini. *Ketiga*, memperluas integrasi data ke kitab-kitab hadis di luar Kutub al-Tisah, seperti Musnad al-Humaidi, Mustadrak al-Hakim, atau Sunan al-Daraquthni, yang juga dikutip dalam al-Mujam al-Mufahras namun belum termasuk dalam repositori digital saat ini, sehingga model yang dikembangkan tidak hanya komprehensif secara internal, tetapi juga inklusif terhadap keragaman sumber hadis dalam tradisi tekstual Islam. Dengan ketiga langkah ini, transformasi digital leksikon hadis tidak hanya menjadi alat preservasi, tetapi evolusi metodologis dalam studi hadis kontemporer berbasis data dan filologi digital.

## DAFTAR PUSTAKA

Al-Mandzur, I. (1928). *Lisanul 'Arabiy*. Darul Ihya al Turats al Arabiy.

Al-Qazwini, A. I. F. (1999). *Mu'jam maqayis al-lughah*. Dar al-Jil. https://books.google.co.id/books?id=zqVDAQAACAAJ

Aziz, A., Sebgag, S., Zuana, M. M. M., & Suryani, I. (2022). Learning Arabic Pegon for Non-Javanese Santri At Pesantren. *Jurnal Pendidikan Islam*, *8*(2), 113–126. https://doi.org/10.15575/jpi.v8i2.19581

Baalbaki, R. (2014). The Arabic Lexicographical Tradition. In *The Arabic Lexicographical Tradition*. https://doi.org/10.1163/9789004274013

Brown, J. (2018). *Hadith : Muhammad's legacy in the medieval and modern world*. Oneworld Academic.

Bunt, G. R. (2018). Conclusion: In *Hashtag Islam* (pp. 141–150). University of North Carolina Press. http://www.jstor.org/stable/10.5149/9781469643182_bunt.12

Cachia, P., Wehr, H., & Cowan, J. M. (1985). A Dictionary of Modern Written Arabic. *Journal of the American Oriental Society*. https://doi.org/10.2307/602745

Dalimunthe, R. P., & Siti, N. (2021). Kontektualisasi Hadis : Menyikapi penomena prank di Media Sosial. *Diroyah Jurnal Studi Islam*, *5*(2).

Driscoll, M. J., & Pierazzo, E. (2016). Digital scholarly editing: Theories and practices. In *Digital Scholarly Editing: Theories and Practices*. https://doi.org/10.11647/OBP.0095

Fauzi, I. (2020). HADIS DARI KLASIK LITERAL KE PORTABLE DIGITAL: Telaah Aplikasi Smartphone Mausu'ah al-Hadis al-Syarif Islamweb. *Riwayah : Jurnal Studi Hadis*, *6*(1), 1. https://doi.org/10.21043/riwayah.v6i1.6747

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic *Natural Language Processing*: An overview. *Journal of King Saud University - Computer and Information Sciences*, *33*(5), 497–507. https://doi.org/10.1016/j.jksuci.2019.02.006

Haywood, J., Wehr, H., & Cowan, J. M. (1980). A Dictionary of Modern Written Arabic (Arabic-English). *Die Welt Des Islams*. https://doi.org/10.2307/1569532

Hitti, P. K., Wensinck, A. J., Grousset, R., Munro, D. C., Halkin, A. S., Ivanow, W., Tusi, N., Ivanow, W., din Shah, S., Ivanow, Maḥmūd, 'Abbās, Ḥimādeh, S. B., Chol, I. B., Zurayq, C. K., al-Maqdisi, A. K., Jabbūr, J. S., al-Shihābi, A. Ḥaydar, Rustum, A., al-Bustāni, F. I., … al-Bustani, F. I. (1936). Concordance et indices de la tradition musulmane. *Journal of the American Oriental Society*. https://doi.org/10.2307/594281

Iryani, J., Hafid, E., & Ahmad, A. (2023). Media Pembelajaran dalam Presfektif Hadis. *PIJAR: Jurnal Pendidikan Dan Pengajaran*, *1*(2), 225–234. https://doi.org/10.58540/pijar.v1i2.216

Kamran, A. B., Abro, B., & Basharat, A. (2023). SemanticHadith: An ontology-driven knowledge graph for the hadith corpus. *Journal of Web Semantics*, *78*, 100797. https://doi.org/10.1016/j.websem.2023.100797

Kamran, A. B., Butt, N. A., & Basharat, A. (2024). Semantic Enrichment of Hadith Corpus - Knowledge Graph Generation from Islamic Text. *Semantic Web Journal*, *0*. https://www.semantic-web-journal.net/system/files/swj3651.pdf

Lehnert, W. G. (1992). *NLP and text analysis at the University of Massachusetts*. https://doi.org/10.3115/1075527.1075668

Najiyah, N. L. N. N., & Putriani, R. (2024). Transformation of Hadith Study in the Digital Era: an Effectiveness of Hadith Applications and Websites. *Mashdar: Jurnal Studi Al-Qur'an Dan Hadis*, *6*(1), 27–42. https://doi.org/10.15548/mashdar.v6i1.7882

Pinto, N., Idris, M., & Sarwan, S. (2022). Hadis dan Media Abad Ke-20 (Penolakan Hadis Dhaif tentang Larangan Wanita Diberi Pendidikan dalam Majalah al-Munir). *Jurnal Ulunnuha*, *11*(2), 168–177. https://doi.org/doi.org/10.15548/ju.v11i2.5539

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

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. https://doi.org/10.1016/j.jksuci.2021.01.019

Siddiqi, M. Z. (1961). *Hadith Literature Its Origin, Development, Special Features and Criticism*. Sibendranath Kanjilal.

Suhendra, A. (2019). Transmisi Keilmuan Pada Era Milenial Melalui Tradisi Sanadan Di Pondok Pesantren Al-Hasaniyah. *Jurnal SMART (Studi Masyarakat, Religi, Dan ….* https://journal.blasemarang.id/index.php/smart/article/view/859

Wahid, A., & Wahyuni, F. S. (2018). *Efektifitas Pembelajaran Hadith Tematik dengan Software dan Aplikasi Shamela Library di Madrash ALiayah An-Nur Al-Huda Ngawonggo Tajinan Malang*. *1*(1), 43–49. https://doi.org/10.36040/mnemonic.v1i1.19

Wahyuningsih, S., & Istianah, I. (n.d.). Kontribusi Digitalisasi Hadis Bagi Perkembangan Studi Hadis di Era Revolusi Industri 4.0. In *repository.iainkudus.ac.id*. http://repository.iainkudus.ac.id/7361/1/Buku Digitalisasi .pdf

Xu, D., Jansen, P., Martin, J., Xie, Z., Yadav, V., Madabushi, H. T., Tafjord, O., & Clark, P. (2020). Multi-class hierarchical question classification for multiple choice science exams. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, *May*, 5370–5382.

Yeni, F., Pinto, N., & Gonsales, G. (2024). Living Hadith Studies : A VOSviewer Perspective. *Mashdar: Jurnal Studi Al-Qur'an Dan Hadis*, *6*(2), 173–186. https://doi.org/10.15548/mashdar.v6i2.9514