TESIS

MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF

Disusun oleh:

AHMAD HASYIM ASY'ARI

23.55.2522

Konsentrasi : Business Intelligence

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

TESIS

MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF

HADITH TOPIC CLUSTERING MODEL IN BUKHARI-MUSLIM HADITH BASED ON BERT EMBEDDING INTEGRATION WITH ADDITIONAL SEMANTIC FEATURES OF TEXT LENGTH AND TF-IDF

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi S2 PJJ Teknik Informatika

Disusun oleh:

Nama : **Ahmad Hasyim Asy'ari**
NIM : **23.55.2522**
Konsentrasi : **Business Intelligence**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF

HADITH TOPIC CLUSTERING MODEL IN BUKHARI-MUSLIM HADITH BASED ON BERT EMBEDDING INTEGRATION WITH ADDITIONAL SEMANTIC FEATURES OF TEXT LENGTH AND TF-IDF

yang disusun dan diajukan oleh

Ahmad Hasyim Asy'ari

23.55.2522

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 10 Oktober 2025

Dosen Pembimbing,

Hanafi, S.Kom., M.Eng., Ph.D.

NIK. 190302024

HALAMAN PENGESAHAN

MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF

HADITH TOPIC CLUSTERING MODEL IN BUKHARI-MUSLIM HADITH BASED ON BERT EMBEDDING INTEGRATION WITH ADDITIONAL SEMANTIC FEATURES OF TEXT LENGTH AND TF-IDF

yang disusun dan diajukan oleh

Ahmad Hasyim Asy'ari

23.55.2522

Telah dipertahankan di depan Dewan Penguji

pada tanggal 10 Oktober 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Hanafi, S.Kom., M.Eng., Ph.D. **NIK. 190302024**	
Dhani Ariatmanto, S.Kom., M.Kom., Ph.D. **NIK. 190302197**	
Alva Hendi Muhammad, S.T., M.Eng., Ph.D. **NIK. 190302493**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 10 Oktober 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ahmad Hasyim Asy'ari
NIM : 23.55.25.22

Menyatakan bahwa Tesis dengan judul berikut:

Model Klasterisasi Topik Hadis Pada Hadis Bukhari-Muslim Berbasis Integrasi Embedding Bert Dengan Fitur Semantik Tambahan Panjang Teks dan TF-IDF

Dosen Pembimbing : Hanafi, S.Kom., M.Eng., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 10 Oktober 2025

Yang Menyatakan,

Ahmad Hasyim Asy'ari

HALAMAN PERSEMBAHAN

Segala puji dan syukur saya panjatkan ke hadirat Allah SWT atas limpahan rahmat, hidayah, dan karunia-Nya yang tak terhingga. Sehingga penulis dapat merampungkan tesis ini yang berjudul "MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF".

Penelitian ini kami persembahkan kepada kedua orang tua dan istri tercinta yang selalu menjadi sumber kekuatan dalam setiap langkah sehingga bisa menyelesaikan studi Program Magister (S2) pada Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI) Universitas Amikom Yogyakarta.

HALAMAN MOTTO

"Barangsiapa menempuh jalan untuk mencari ilmu, maka Allah akan memudahkan baginya jalan menuju surga." *(HR. Muslim).*

"Sampaikanlah dariku walau hanya satu ayat." *(HR. Bukhari).*

"Ilmu lebih baik daripada harta. Ilmu menjaga engkau, sedangkan harta engkau yang menjaganya." *(Ali bin Abi Thalib RA).*

"Data tanpa makna adalah kebisingan. Hadis tanpa pemahaman adalah kehilangan. Ilmu hadir untuk menjembatani keduanya."

"Mengelompokkan hadis dalam topik-topik bukan hanya soal ilmu data, tapi juga bentuk cinta kepada sabda Nabi."

KATA PENGANTAR

Segala puji dan syukur saya panjatkan ke hadirat Allah SWT atas limpahan rahmat, hidayah, dan karunia-Nya yang tak terhingga. Shalawat serta salam semoga senantiasa tercurah kepada junjungan kita Nabi Muhammad SAW, sebagai suri teladan sepanjang zaman, yang telah membawa umat manusia dari zaman kegelapan menuju cahaya ilmu dan kebenaran.

Dengan pertolongan Allah SWT, penulis dapat menyelesaikan tesis ini yang berjudul "MODEL KLASTERISASI TOPIK HADIS PADA HADIS BUKHARI-MUSLIM BERBASIS INTEGRASI EMBEDDING BERT DENGAN FITUR SEMANTIK TAMBAHAN PANJANG TEKS DAN TF-IDF".

Terimakasih atas segala bentuk dukungan lahir dan batin, sehingga bisa menyelesaikan studi Program Magister (S2) pada Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI) Universitas Amikom Yogyakarta. Untuk itu penulis mempersembahkan tesis ini dengan sepenuh hati kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas Amikom Yogyakarta.
2. Ibu Prof. Dr. Kusrini, M.Kom., selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta yang telah memberikan kesempatan dan izin untuk menempuh studi lanjut di Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI).
3. Bapak Hanafi, S.Kom., M.Eng., Ph.D., selaku pembimbing utama. Yang bersedia menyempatkan waktu di sela-sela kesibukan beliau, untuk

memberikan dukungan, membimbing, mengoreksi dan mengarahkan penulis demi kesempurnaan penulisan penelitian ini. . Semoga ilmu dan kebaikan yang diberikan menjadi amal yang terus mengalir.

4. Tim Penguji dari SPT, SHPT, hingga UT yang telah memberikan arahan dan wawasan lebih dalam proses penyempurnaan penulisan.
5. Seluruh Dosen Pengajar di Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI) Universitas Amikom Yogyakarta dari semester pertama hingga terakhir yang memberikan arahan, dukungan, semangat, dan sharing pengetahuan sehingga penulis mendapatkan wawasan baru yang lebih luas dalam menyelesaikan tugas disetiap studi.
6. Segenap Civitas Akademika (Pengelola dan Admisi) Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI) Universitas Amikom Yogyakarta yang telah memberikan pelayanan dan bantuan sangat baik dalam kebutuhan studi.
7. Kedua orang tua Alm. H. Zaenal hamami, Almh. Kholidiyah dan istri tercinta Dr.Kharisma Eka Putri, M.Pd. yang selalu menjadi sumber kekuatan dalam setiap langkah hidup saya. Terima kasih atas kasih sayang, doa yang tak pernah putus, dan dukungan lahir batin yang tak ternilai.
8. Keluarga besar yang di rumah mauapun yang di SMAN 1 Purwoasri yang senantiasa memberikan semangat dan kepercayaan, menjadi tempat kembali dan berbagi di setiap suka dan duka selama masa studi ini.

9. Teman-teman Mahasiswa Program Studi Pendidikan Jarak Jauh Magister Teknik Informatika (PJJ-MTI) Universitas Amikom Yogyakarta Angkatan 2023.

Dengan senang hati penulis menerima kritik dan saran yang membangun dari pembaca. Karena penulis menyadari dengan penuh bahwa dalam penyusunan penelitian ini masih banyak kekurangan.

Akhir kata, semoga penelitian ini dapat memberikan manfaat bagi pembacanya dan menjadi amal kebaikan di hadapan Allah SWT, serta sejalan dengan harapan dan niat yang tulus.

Yogyakarta, 10 Oktober 2025

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Klastering hadis merupakan tugas penting dalam studi Islam, mengingat sifat korpus hadis yang luas dan kompleks. Pendekatan pengelompokan tradisional sering kali kesulitan untuk menangkap konteks semantik yang mendalam dalam hadis, yang menyebabkan pengelompokan topik menjadi kurang akurat. Kemajuan terkini dalam Natural Language Processing (NLP), seperti model Bidirectional Encoder Representations from Transformers (BERT), telah menunjukkan hasil yang menjanjikan dalam mengatasi tantangan ini dengan menyediakan penyematan kontekstual yang kaya. Namun, penggunaan BERT secara tunggal dapat mengabaikan fitur linguistik yang penting, yang berpotensi membatasi kinerja pengelompokan. Studi ini mengusulkan model pengelompokan yang disempurnakan untuk koleksi hadis Sahih Bukhari dan Sahih Muslim, yang mengintegrasikan penyematan BERT dengan fitur semantik tambahan, termasuk panjang teks, Term Frequency (TF), dan Inverse Document Frequency (IDF). Dengan menggunakan kerangka BERTopic, pendekatan ini menangkap hubungan yang bernuansa antara hadis, yang memberikan hasil pengelompokan yang lebih akurat secara kontekstual. Eksperimen menunjukkan bahwa metode terintegrasi ini secara signifikan meningkatkan kinerja pengelompokan, seperti yang ditunjukkan oleh silhouette score dengan nilai -0.1 dan davies-bouldin index 2.6. Sedangkan tanpa terintegrasi menunjukkan nilai rendah dengan silhouette score dengan nilai -0.145 dan davies-bouldin index 6.6. Sehingga pengembangan ini menawarkan metode yang lebih tepat untuk pengelompokan topik dalam studi Islam, yang memfasilitasi organisasi dan pemahaman yang lebih baik tentang teks hadis.

Kata kunci: Klasterisasi Hadis, Fitur Semantik, BERTopic, NLP

ABSTRACT

Hadith clustering is an important task in Islamic studies, given the vast and complex nature of the hadith corpus. Traditional clustering approaches often struggle to capture the deep semantic context in hadith, leading to inaccurate topic clustering. Recent advances in Natural Language Processing (NLP), such as the Bidirectional Encoder Representations from Transformers (BERT) model, have shown promise in addressing this challenge by providing rich contextual embeddings. However, using BERT alone may overlook important linguistic features, potentially limiting clustering performance. This study proposes an enhanced clustering model for Sahih Bukhari and Sahih Muslim hadith collections, integrating BERT embeddings with additional semantic features, including text length, term frequency (TF), and inverse document frequency (IDF).Using the BERTopic framework, this approach captures the nuanced relationships between hadiths, providing a more contextually accurate clustering output. Experiments show that this integrated method significantly improves clustering performance, as indicated by the silhouette score with a value of -0.1 and davies-bouldin Index of 2.6. While without integration shows a low value with a silhouette score with a value of -0.145 and davies-bouldin index 6.6. So that, this development offers a more appropriate method for topic clustering in Islamic studies, which facilitates better organization and understanding of hadith texts.

Keyword: *Hadis, Semantic Features, BERTopic, NLP*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Sahih Bukhari, yang disusun oleh Imam Muhammad ibn Ismail al-Bukhari pada abad ke-9 M, adalah salah satu koleksi hadis yang paling dihormati dalam Islam Sunni. Ia diakui karena metodologinya yang ketat dalam mengotentikasi hadis, yang merupakan ucapan dan tindakan Nabi Muhammad SAW. Pentingnya Sahih Bukhari terletak tidak hanya pada isinya tetapi juga dalam perannya dalam membentuk yurisprudensi Islam dan bimbingan moral. Koleksi ini terdiri dari lebih dari 7.000 hadis, yang telah dikategorikan dan diautentikasi dengan cermat, menjadikannya landasan pengetahuan Islam [1] [2].

Studi hadis, termasuk Sahih Bukhari, telah berkembang selama berabad-abad, yang mencerminkan sifat dinamis dari pemikiran dan praktik Islam. Periode Abbasiyah, misalnya, menandai perkembangan signifikan dalam literatur hadis, karena para ulama berusaha untuk melestarikan dan mengotentikasi ajaran Nabi di tengah berbagai perubahan sosial-politik [3]. Era ini menyaksikan pembentukan metodologi formal untuk klasifikasi dan otentikasi hadis, yang masih relevan hingga saat ini. Ulama seperti al-Bukhari menggunakan kriteria yang ketat untuk menerima hadis, menekankan keandalan perawi dan kontinuitas transmisi [4][5].

Di zaman kontemporer, relevansi studi hadis telah meluas ke berbagai bidang, termasuk studi gender dan humaniora digital. Misalnya, penelitian terkini telah mengeksplorasi penafsiran ulang hadis terkait gender dalam konteks sosial

budaya modern, menekankan peran perempuan dalam sejarah Islam awal dan kontribusinya terhadap masyarakat [6][7][8]. Lebih jauh lagi, munculnya teknologi telah mendorong para cendekiawan untuk menggunakan teknik pembelajaran mesin untuk klasifikasi dan autentikasi hadis, yang menyoroti persinggungan antara keilmuan tradisional dan metodologi modern [1][9].

Dengan banyaknya penelitian terkini mengenai Sahih Bukhari dan bidang studi hadis yang lebih luas merupakan cerminan kekayaan ilmu pengetahuan Islam yang telah berkembang selama berabad-abad. Metodologi ketat yang ditetapkan oleh para ulama terdahulu, dikombinasikan dengan penafsiran ulang kontemporer dan kemajuan teknologi, memastikan bahwa hadis tetap menjadi sumber petunjuk penting bagi umat Islam saat ini.

Studi tentang *Natural Language Processing* (NLP) yang diterapkan pada literatur Hadis, khususnya Sahih Bukhari dan Sahih Muslim, telah mendapatkan daya tarik dalam beberapa tahun terakhir karena meningkatnya kebutuhan akan metode otomatis untuk menganalisis dan mengklasifikasikan teks-teks penting ini dalam kajian ilmiyah Islam. Sahih Bukhari, yang disusun oleh Imam Muhammad ibn Ismail al-Bukhari, adalah salah satu koleksi Hadis yang paling autentik, dan dokumentasinya yang cermat tentang perkataan dan tindakan Nabi Muhammad SAW. menjadikannya subjek penting untuk aplikasi NLP [7]. Integrasi teknik NLP memfasilitasi ekstraksi wawasan yang bermakna dari korpus Hadis yang luas, yang secara tradisional tidak terstruktur dan menantang untuk dinavigasi.

Perkembangan NLP dalam beberapa tahun terakhir telah menghadirkan metode representasi teks yang semakin canggih. Salah satu metode yang banyak

digunakan dalam penggalian topik adalah *Latent Dirichlet Allocation* (LDA) yang mengandalkan pendekatan probabilistik berbasis frekuensi kata. Meskipun populer karena kesederhanaan dan interpretabilitasnya, LDA memiliki sejumlah keterbatasan. Model ini hanya memandang teks sebagai kumpulan kata (*bag-of-words*) tanpa memperhatikan urutan dan konteks, sehingga gagal menangkap makna semantik yang lebih dalam. Selain itu, performa LDA menurun pada teks pendek atau teks dengan morfologi kompleks, serta sulit membedakan sinonimi maupun polisemik yang sering muncul dalam bahasa alami[10][11].

Sebaliknya, hadirnya Bidirectional Encoder Representations from Transformers (BERT) membawa terobosan dalam pemrosesan teks. BERT menggunakan arsitektur transformer encoder yang membaca teks secara dua arah, dari kiri ke kanan dan kanan ke kiri sekaligus, sehingga mampu memahami makna kata dalam konteks kalimat penuh [12]. Representasi yang dihasilkan berupa dense embeddings yang kaya makna semantik dan dapat diaplikasikan pada berbagai tugas, termasuk klasterisasi topik. Penelitian mutakhir menunjukkan bahwa integrasi BERT dengan metode klasterisasi mampu menghasilkan topik yang lebih koheren dan mudah diinterpretasi dibandingkan dengan LDA maupun model klasik lainnya [13] [14].

Keunggulan lain BERT terletak pada kemampuannya dalam menangani teks pendek maupun domain khusus. Sementara LDA sering gagal mengidentifikasi topik dari teks singkat, embedding BERT tetap dapat merepresentasikan makna dengan baik. Hal ini terbukti dari berbagai studi yang menunjukkan peningkatan performa klasterisasi ketika menggunakan BERT atau turunannya seperti Sentence-

BERT [15]. Selain itu, fleksibilitas BERT memungkinkan representasi teks yang sama digunakan kembali untuk berbagai downstream task seperti klasifikasi, pencarian semantik, maupun ekstraksi entitas, sehingga pipeline penelitian menjadi lebih serbaguna [16].

Dalam konteks penelitian hadis, khususnya hadis Shahih Bukhari-Muslim yang memiliki variasi redaksi, istilah domain-spesifik, serta struktur sanad dan matan yang kompleks, penggunaan BERT menjadi sangat relevan. Representasi berbasis semantik dari BERT memungkinkan pemetaan topik hadis secara lebih akurat tanpa bergantung pada label manual yang sulit dan mahal diperoleh. Dengan demikian, pemilihan BERT dalam penelitian "Model Klasterisasi Topik Hadis pada Hadis Shahih Bukhari Muslim" diharapkan mampu mengatasi keterbatasan metode klasik seperti LDA sekaligus menghasilkan topik yang lebih koheren, bermakna, dan bermanfaat untuk pengembangan kajian keilmuan Islam di era digital [11] [14].

Perkembangan pesat Natural Language Processing (NLP) mendorong pemanfaatan representasi berbasis transformer seperti BERT untuk mengekstraksi makna semantik teks yang kaya konteks. Dalam penggalian topik, dua paradigma pembelajaran mesin paling menonjol adalah klasterisasi (*unsupervised learning*) dan klasifikasi (*supervised learning*). Keduanya berbeda secara tujuan, kebutuhan data, dan keluaran. Klasifikasi memerlukan label untuk memetakan teks ke kelas yang telah ditentukan, sedangkan klasterisasi menemukan struktur laten atau kelompok-kelompok serupa tanpa label awal [17]. Perbedaan ini memiliki implikasi metodologis yang signifikan terutama ketika label berkualitas sulit

diperoleh, seperti pada korpus hadis berbahasa Arab klasik dengan struktur unik isnad dan matan.

Pada ranah pemodelan topik modern, teknik berbasis embedding misalnya BERTopic menggabungkan embedding semantik (BERT/SBERT), reduksi dimensi, dan pengelompokan untuk menghasilkan topik yang lebih tajam dan mudah diinterpretasikan dibanding pendekatan klasik berbasis frekuensi. Secara teknis, BERTopic memanfaatkan representasi kalimat padat dan skema c-TF-IDF untuk merangkum istilah kunci per topik, sehingga cocok untuk korpus besar dan heterogen [13] [14][16].

Konteks hadis khususnya Shahih Bukhari Muslim menyajikan tantangan linguistik dan struktural variasi redaksi, rujukan sanad berlapis, serta istilah domain-spesifik. Riset NLP terkini di studi keislaman menunjukkan ketersediaan dataset dan sumber baru, seperti Multi-IsnadSet untuk Shahih Muslim[18], serta upaya topic extraction yang secara eksplisit memanfaatkan materi hadis sebagai korpus penelitian [19]. Hal ini membuka peluang untuk pendekatan unsupervised topic clustering yang mampu memetakan tema-tema fikih, akidah, adab, hingga konteks perawi tanpa ketergantungan pada label kuratorial yang mahal.

Di sisi algoritmik, pemilihan teknik klasterisasi memengaruhi kualitas topik. K-Means menonjol karena kesederhanaan dan skalabilitas, namun mengasumsikan bentuk klaster cenderung bulat dan membutuhkan jumlah klaster k yang ditentukan di awal. Tinjauan mutakhir menyoroti banyak varian dan optimisasi K-Means untuk meningkatkan robustnes serta efisiensi pada data berdimensi tinggi seperti embedding [20]. Sebaliknya, HDBSCAN mampu

menemukan klaster dengan bentuk tak beraturan dan mengidentifikasi noise/outlier tanpa harus menentukan k, sehingga kerap lebih stabil pada embedding BERT/SBERT. Studi terkini tentang klasterisasi teks berbasis SBERT juga menegaskan pentingnya pilihan teknik pooling dan penanganan teks panjang agar struktur topik lebih konsisten [15][21].

Dengan mempertimbangkan perbedaan klasterisasi, klasifikasi dan karakteristik korpus hadis, pendekatan klasterisasi topik berbasis BERT menjadi relevan untuk eksplorasi ilmu memetakan tema-tema hadis tanpa bias label awal, navigasi pengetahuan membantu peneliti/mahasiswa menelusuri tema lintas kitab dan perawi, dasar anotasi lanjut menyediakan kerangka awal yang kemudian dapat dipakai sebagai silver labels untuk rancangan klasifikasi terarah atau *supervised* di tahap berikutnya [17].

Studi terkini telah menunjukkan kegunaan BERT dalam klasterisasi dokumen dengan mengodekan dokumen ke dalam representasi vektor yang padat. Misalnya, Dodda [22] menyoroti bagaimana penyematan pra-latihan BERT dapat disetel dengan baik untuk tugas klasterisasi tertentu, yang memungkinkan adaptasi terhadap karakteristik unik dari kumpulan data yang berbeda. Proses penyempurnaan ini penting, karena meningkatkan kemampuan model untuk menangkap pola semantik yang melekat dalam data, sehingga meningkatkan kinerja klasterisasi.

Selain itu, integrasi BERT dengan algoritma klasterisasi tradisional telah menunjukkan hasil yang menjanjikan. Gupta dan Gupta [23] mengusulkan penanganan disambiguasi kata sebagai masalah klasterisasi lokal, dengan

memanfaatkan vektor kata kontekstual yang berasal dari BERT. Pendekatan ini tidak hanya meningkatkan kualitas representasi dokumen tetapi juga memfasilitasi hasil klasterisasi yang lebih efektif. Demikian pula, penelitian seperti yang dilakukan oleh [24] telah mengeksplorasi berbagai algoritma klasterisasi, termasuk k-means dan fuzzy c-means, bersama dengan representasi BERT, yang selanjutnya memvalidasi fleksibilitas model dalam aplikasi klasterisasi.

Selain metode klasterisasi tradisional, pendekatan hibrida yang menggabungkan BERT dengan teknik pemodelan topik telah diteliti. George dan Sumathy [25] mengusulkan kerangka kerja terpadu yang mengintegrasikan BERT dengan *Latent Dirichlet Allocation* (LDA) untuk pemodelan topik yang lebih baik. Temuan mereka menunjukkan bahwa klasterisasi dapat meningkatkan koherensi topik yang diekstrak dari korpus teks besar, yang menunjukkan potensi penggabungan penyematan kontekstual BERT dengan metodologi klasterisasi untuk menghasilkan wawasan yang lebih bermakna dari data teks.

Gambar 1.1. Hasil Perbandingan LDA, BERT, LDA-BERT

Hasil eksperimen George dan Sumathy [25] menunjukkan bahwa pendekatan yang diusulkan berkinerja lebih baik daripada algoritma sebelumnya

yang dilaporkan dalam penelitian serupa. Hasilnya dibandingkan dengan algoritma LDA dan BERT yang sudah ada. Gambar 1.1 menunjukkan bahwa pengelompokan terintegrasi dan kerangka kerja LDA-BERT gabungan yang diusulkan berkinerja lebih baik daripada metode lain yang dipertimbangkan. Dan juga metode reduksi dimensionalitas UMAP dilakukan di atas LDA-BERT gabungan untuk mendapatkan fitur berbasis konteks yang lebih presisi guna meningkatkan pemodelan topik.

Dengan sumber berbagai hasil penelitian sebelumnya dan seperti pada tabel hasil penelitian Arun Kumar Yadaf dll. [26] menerangkan bahwa algoritma BERT telah memajukan bidang klasterisasi dokumen secara signifikan dengan menyediakan penyematan kontekstual yang kuat yang meningkatkan pemahaman semantik teks. Kombinasi BERT dengan berbagai teknik klasterisasi, termasuk algoritma tradisional dan kerangka kerja pemodelan topik atau Hybrid Model, telah terbukti efektif dalam mengekstraksi pola yang bermakna dari berbagai kumpulan data. Ini dibuktikan pada Tabel 1.1 Perbedaan Aspek terhadap Topik model LDA, BERT dan BERT-LDA bahwa pemodelan topik atau Hybrid Model memiliki peforma yang tinggi pada aspek hubungan kontekstual dan hubungan topik. Arah penelitian di masa mendatang dapat difokuskan pada penyempurnaan lebih lanjut pendekatan ini dan mengeksplorasi integrasi model yang lebih canggih untuk meningkatkan hasil klasterisasi.

Tabel 1.1. Perbedaan Aspek terhadap Topik model LDA, BERT dan BERT-LDA

Aspek	*Traditional LDA*	*BERT Only*	*Proposed Work A Hybrid Model Integrating LDA, BERT*
Contextual Relationship	*Low*	*High*	*High*
Topic Coherence	*Moderate*	*High*	*High*
Scalability	*High*	*Moderate*	*Moderate*
Computational Cost	*Low*	*High*	*Moderate to High*

Penerapan algoritma BERT untuk klasterisasi topik-topik hadis, khususnya dari koleksi Sahih al-Bukhari dan Sahih Muslim, merupakan kemajuan signifikan dalam bidang studi Islam dan pemrosesan bahasa alami NLP. Kemampuan BERT untuk menghasilkan penyematan kontekstual memungkinkan pemahaman yang bernuansa tentang hubungan semantik dalam teks-teks hadis, yang sangat penting untuk klasterisasi yang efektif.

Penyematan kontekstual BERT dapat digunakan untuk mengubah teks-teks hadis menjadi representasi vektor yang padat, menangkap makna dan hubungan yang rumit di antara hadis yang berbeda. Transformasi ini penting untuk klasterisasi, karena memungkinkan klasterisasi hadis berdasarkan konten semantiknya. Abdelaal dkk. [1] menyoroti efektivitas algoritma pembelajaran terbimbing dalam mengklasifikasikan hadis ke dalam kategori-kategori yang berbeda, yang dapat berfungsi sebagai pendahulu klasterisasi. Studi mereka menunjukkan bahwa BERT dapat secara signifikan meningkatkan akurasi klasifikasi hadis, menjadikannya kandidat yang cocok untuk aplikasi klasterisasi.

Selain itu, mengintegrasikan BERT dengan teknik pemodelan topik dapat lebih meningkatkan klasterisasi topik Hadis. Peinelt [27] menyarankan bahwa menggabungkan BERT dengan model topik tradisional dapat meningkatkan deteksi

kesamaan semantik, yang penting untuk tugas klasterisasi. Pendekatan ini dapat membantu dalam mengidentifikasi tema-tema yang mendasari dalam korpus Hadis, memfasilitasi pemahaman yang lebih terstruktur tentang topik-topik yang dibahas dalam Sahih al-Bukhari dan Sahih Muslim. Penelitian penggabungan BERT [26]dengan model topik tradisional (LDA) juga memberikan hasil skor siluet yang baik dengan ditunjukkan pada tabel 1.2 di bawah ini.

Tabel 1.2 *Silhouette Score* yang diperoleh untuk berbagai model menggunakan metode pengurangan dimensi yang berbeda

Topic Models	PCA	t-SNE	UMAP
LDA	0.33239	0.346303	0.37624
BERT	0.46211	0.36938	0.48761
BERT-LDA	0.50843	0.471261	0.51998

Dalam pemodelan topik pada Tabel 1.2. metode tradisional seperti Latent Dirichlet Allocation (LDA) memiliki keterbatasan dalam memahami konteks semantik sehingga menghasilkan klaster yang kurang koheren, sebagaimana terlihat dari nilai *Silhouette Score* yang rendah, yaitu 0.33239 (PCA), 0.346303 (t-SNE), dan 0.37624 (UMAP). Kemudian, penggunaan Bidirectional Encoder Representations from Transformers (BERT) meningkatkan kualitas representasi teks dengan nilai *Silhouette Score* yang lebih tinggi, yakni 0.46211, 0.369380, dan 0.48761. Namun, pendekatan ini masih kurang dalam interpretasi topik secara eksplisit. Untuk mengatasi hal tersebut, model gabungan BERT-LDA dikembangkan dengan mengintegrasikan keunggulan semantik BERT dan struktur probabilistik LDA. Hasilnya menunjukkan peningkatan signifikan dengan Silhouette Score tertinggi, yaitu 0.50843 (PCA), 0.471261 (t-SNE), dan 0.51998 (UMAP), yang menandakan bahwa model ini mampu menghasilkan klaster topik

yang lebih jelas, koheren, dan bermakna secara semantik. Oleh karena itu, kombinasi BERT dan LDA dianggap sebagai pendekatan yang lebih unggul untuk menghasilkan model topik yang tidak hanya semantik tetapi juga interpretatif, khususnya dalam konteks analisis teks keagamaan seperti hadis yang memiliki kompleksitas makna dan konteks yang tinggi.

Dengan pemaparan di atas maka algoritma BERT menawarkan kerangka kerja yang kuat untuk mengelompokkan topik hadis dari Sahih al-Bukhari dan Sahih Muslim dengan memanfaatkan penyematan kontekstual dan kemampuan beradaptasi terhadap berbagai metodologi klasterisasi. Kombinasi BERT dengan teknik pembelajaran terbimbing, pemodelan topik, dan grafik pengetahuan dapat menghasilkan klasterisasi hadis yang lebih bermakna berdasarkan isinya. Secara praktis, penelitian ini memfokuskan pada perumusan pipeline BERT → reduksi dimensi → klasterisasi (K-Means dengan HDBSCAN) → interpretasi topik (c-TF-IDF) pada korpus hadis Shahih Bukhari–Muslim. Perbandingan hasil antar algoritme diharapkan memberi Gambaran empiris mengenai koherensi topik, kestabilan klaster, dan keterjelasan label topik pada teks keagamaan klasik, sekaligus memperkaya literatur NLP keislaman kontemporer [16][14]. Penelitian di masa mendatang dapat difokuskan pada pengoptimalan pendekatan ini dan mengeksplorasi integrasi informasi kontekstual tambahan untuk meningkatkan hasil klasterisasi.

1.2. Rumusan Masalah

Berdasarkan penjelasan pada latar belakang masalah, maka rumusan masalah pada penelitian ini adalah sebagai berikut :

a. Bagaimana cara model algoritma BERT dapat diterapkan secara efektif untuk klasterisasi hadis dari Sahih Bukhari dan Muslim?
b. Topik apa saja yang akan dominan pada klasterisasi hadis dari Sahih Bukhari dan Muslim?
c. Bagaimana akurasi pada penerapan model algoritma BERT untuk klasterisasi hadis dari Sahih Bukhari dan Muslim?

1.3. Batasan Masalah

Berikut batasan-batasan masalah pada penelitian ini:

a. Data yang dianalisa adalah dataset hadis Sohih Bukhori dan Muslim yang diperoleh dari (https://www.kaggle.com/datasets).
b. Model yang digunakan untuk klasterisasi adalah model Algoritma BERT tidak berlabel.
c. Penelitian ini tidak akan mencakup pengujian di luar hadis Bukhori Muslim.

1.4. Tujuan Penelitian

Bagian ini memuat penjelasan secara spesifik:

a. Mengembangkan model algoritma BERT untuk klasterisasi hadis dari Sahih Bukhari dan Muslim.

b. Mendapatkan Gambaran topik yang dominan pada klasterisasi hadis dari Sahih Bukhari dan Muslim.

c. Menguji akurasi pada penerapan model algoritma BERT untuk klasterisasi hadis dari Sahih Bukhari dan Muslim.

1.5. Manfaat Penelitian

Manfaat yang dapat diambil pada penelitian ini adalah sebagai berikut:

a. Hasil dari penelitian ini dapat menjadi dasar atau landasan untuk penelitian lebih lanjut yang mungkin ingin mengembangkan atau memperbaiki metode yang diusulkan.

b. Penelitian ini dapat menjadi rujukan bagi penelitian selanjutnya yang ingin membandingkan metode algoritma BERT untuk klasterisasi hadis dari Sahih Bukhari dan Muslim pada situasi yang lebih kompleks.

c. Dapat menjadi titik awal untuk penelitian lebih mendalam dalam mengoptimalkan parameter, arsitektur, atau teknik lainnya yang digunakan.

1.6. Hipotesis

Model klasterisasi topik Hadis yang diusulkan dalam koleksi hadis Bukhari dan Muslim menggunakan algoritma BERT akan menjadikan hasil akurasi dan relevansi yang baik pada teknik NLP untuk klasterisasi dan analisis teks Hadis. Integrasi model dan metodologi tingkat lanjut tidak hanya akan meningkatkan akurasi klasterisasi topik tetapi juga berkontribusi pada bidang studi Islam yang lebih luas dengan memberikan wawasan yang lebih dalam tentang elemen tematik literatur Hadis.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Topik penelitian pengelompokan topik hadis yang memanfaatkan algoritma BERT, khususnya yang berfokus pada hadis Sahih Bukhari dan Sahih Muslim, merupakan persimpangan antara pemrosesan bahasa alami NLP dan studi Islam. Penerapan BERT, model transformator yang telah dilatih sebelumnya, dalam konteks ini sangat penting untuk memahami dan mengkategorikan korpus literatur hadis yang luas. Sintesis literatur relevan berikut memberikan Gambaran umum yang komprehensif tentang metodologi, tantangan, dan kemajuan dalam bidang ini.

Algoritma BERT telah merevolusi NLP dengan memungkinkan model untuk memahami konteks dan semantik lebih baik daripada algoritma sebelumnya. Sun et al. membahas pergeseran paradigma dalam NLP, dengan menekankan bagaimana model yang telah dilatih sebelumnya seperti BERT telah mengubah berbagai tugas, termasuk klasifikasi dan pengenalan entitas, yang sangat penting untuk memproses teks hadis secara efektif [28]. Pergeseran ini khususnya relevan dalam konteks studi hadis, di mana klasifikasi teks yang akurat berdasarkan konten tematik sangat penting bagi para sarjana dan praktisi.

Penelitian oleh Ismail juga berkontribusi pada pemahaman representasi ontologi dalam korpus hadis [29]. Penelitian ini penting karena meletakkan dasar untuk mengembangkan kerangka kerja terstruktur yang dapat diintegrasikan dengan teknik NLP seperti BERT untuk manajemen dan pengambilan data yang

lebih efektif dalam studi hadis. Pembentukan ontologi dapat memfasilitasi pengelompokan dan kategorisasi hadis yang lebih baik berdasarkan relevansi tematik.

Selain itu, eksplorasi ontologi Arab untuk teks hadis oleh Muhammed menekankan karakteristik unik literatur hadis dan perlunya pendekatan yang disesuaikan dalam aplikasi NLP [30]. Pengembangan ontologi spesifik dapat meningkatkan kinerja model BERT dengan memberi mereka pengetahuan terstruktur yang mencerminkan seluk-beluk narasi hadis.

Selain itu, penelitian Najeeb tentang pendekatan hibrida menggunakan Model Markov Tersembunyi untuk memproses Isnad Hadis melengkapi kerangka BERT dengan menyediakan metodologi tambahan untuk menganalisis komponen struktural hadis [31]. Pendekatan hibrida ini dapat diintegrasikan dengan BERT untuk meningkatkan efektivitas keseluruhan model pengelompokan topik, yang memungkinkan pemahaman yang lebih komprehensif tentang teks hadis.

Liu dkk. membahas tantangan yang terkait dengan pengelompokan tanpa pengawasan, khususnya kurangnya data berlabel, yang mempersulit penilaian kualitas pengelompokan [32]. Hal ini khususnya relevan untuk studi Hadis, di mana tidak adanya kategori yang telah ditetapkan sebelumnya memerlukan metode evaluasi yang kuat untuk memastikan bahwa pengelompokan mencerminkan pengelompokan tematik yang bermakna. Wawasan dari penelitian ini [32], dapat memandu pemilihan metrik evaluasi yang tepat yang memperhitungkan karakteristik unik teks Hadis.

Watson et al. berkontribusi lebih jauh pada diskusi dengan menyoroti pentingnya metrik kedekatan dalam mengevaluasi kinerja pengelompokan [33]. Karya mereka menunjukkan bahwa metrik yang berbeda dapat menghasilkan hasil yang bervariasi, yang penting saat menerapkan algoritme pengelompokan pada teks Hadis. Pilihan metrik kesamaan dapat memengaruhi seberapa baik algoritma BERT menangkap nuansa narasi Hadis, sehingga memengaruhi hasil pengelompokan secara keseluruhan.

Penelitian mengenai pengklasteran teks menggunakan representasi embedding seperti BERT telah menunjukkan bahwa kombinasi metode berbasis transformer dengan algoritma topik klasik seperti LDA, serta integrasi teknik reduksi dimensi dan clustering, mampu menghasilkan klaster topik yang lebih koheren dan representatif secara semantik. [34] dalam penelitiannya menunjukkan bahwa model hibrida BERT-LDA, jika dikombinasikan dengan teknik reduksi dimensi UMAP dan algoritma k-means clustering, menghasilkan performa topik modeling yang unggul dibandingkan model tunggal. Evaluasi menggunakan metrik Silhouette Score membuktikan bahwa hasil klaster dari pendekatan hibrida tersebut lebih kompak dan terpisah dengan baik. Dan Lebih jauh lagi model BERT-LDA dengan klasterisasi berbasis UMAP-KMeans secara konsisten memberikan skor koherensi (*Coherence Score*) yang lebih tinggi dibandingkan pendekatan konvensional, dengan peningkatan hingga 85% pada beberapa skenario topik [26]. Penerapan model ini pada teks keagamaan seperti hadis membuka potensi besar dalam pemetaan tematik, klasifikasi konten, dan penyajian pengetahuan keislaman secara sistematis dan terstruktur. Hal ini menunjukkan bahwa metode klasterisasi

yang terinformasi dengan vektor embedding serta penguatan konteks semantik melalui BERT mampu mengungkap topik-topik tersembunyi dalam korpus besar secara lebih efektif.

Oleh karena itu, pendekatan serupa dapat diterapkan dalam pengklasteran hadis-hadis Bukhari-Muslim, yang memiliki kompleksitas semantik tinggi, guna menghasilkan pemetaan topik yang akurat dan informatif.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks Literatur Review dan Posisi Penelitian Model Klasterisasi Topik Hadis Pada Hadis Bukhari-Muslim Berbasis Integrasi Embedding BERT Dengan Fitur Semantik Tambahan Panjang Teks Dan TF-IDF

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	BERT Models for Arabic Text Classification: A Systematic Review	Alammary, A. (2022).. Applied Sciences, 12(11), 5720. https://doi.org/10.3390/app12115720	Mengidentifikasi model BERT yang digunakan untuk klasifikasi teks bahasa Arab. Membandingkan kinerjanya. Menilai efektivitasnya dibandingkan dengan model BERT bahasa Inggris asli	Tinjauan ini menganalisis 48 artikel dan mengidentifikasi sembilan model BERT yang berbeda untuk klasifikasi teks bahasa Arab. Model berkinerja tinggi meliputi MARBERT, QARiB, ARBERT, AraBERT, dan ArabicBERT, semuanya dilatih sebelumnya pada korpus bahasa Arab yang besar.	Penelitian ini menyarankan untuk mengeksplorasi area penelitian baru untuk perbaikan lebih lanjut dalam klasifikasi teks bahasa Arab menggunakan model BERT Kelemahan yang dicatat adalah terbatasnya penerapan BERT dalam bahasa yang kaya sumber daya seperti bahasa Inggris, yang menunjukkan perlunya lebih banyak fokus pada bahasa Arab	Perbandingan dengan Klasterisasi Hadis, Penelitian ini tidak secara khusus membahas penerapan pemodelan BERT dalam klasterisasi Hadis, seperti dalam Sahih Bukhari dan Muslim. Namun, metodologi dan temuan mengenai klasifikasi teks Arab dapat memberikan wawasan untuk aplikasi serupa dalam teks-teks keagamaan.
2	ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic	Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. (2021). https://doi.org/10.18653/v1/2021.acl-long.551	Tujuan utama adalah untuk mengembangkan dua model bahasa Transformer khusus bahasa Arab yang canggih, ARBERT dan MARBERT.	Model-model tersebut mencapai hasil mutakhir disebagian besar tugas yang dievaluasi dalam tolok ukur ARLUE yang baru diperkenalkan, mengungguli AraBERT dan XLM-R Large	Penelitian ini menunjukkan bahwa tolok ukur ARLUE dapat memfasilitasi evaluasi dan perbandingan di masa mendatang dalam NLP bahasa Arab	Kinerja ARBERT dan MARBERT berpotensi melampaui model BERT tradisional dalam konteks ini karena pelatihan khusus mereka pada kumpulan data bahasa Arab.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	BERT based Named Entity Recognition for Automated Hadith Narrator Identification	Luthfi, E., Yusoh, Z., & Aboobaider, B. (2022). International Journal of Advanced Computer Science and Applications, 13(1). https://doi.org/10.14569/ijacsa.2022.0130173	Tujuan utama adalah untuk memanfaatkan Pengenalan Entitas Bernama (NER) untuk mengidentifikasi dan mengautentikasi perawi Hadis, dengan pengklasifikasi feed-forward tambahan	Model NER yang diusulkan mencapai skor F1 sebesar 99,63% selama pelatihan dan 98,27% dalam ujian akhir, yang menunjukkan efektivitas tinggi dalam mengidentifikasi perawi Hadis	Penelitian ini menyarankan eksplorasi lebih lanjut tentang pengoptimalan BERT untuk NER dalam teks Hadis, terutama untuk bahasa selain bahasa Indonesia. Namun, penelitian ini mencatat kurangnya penelitian ekstensif yang mengoptimalkan BERT untuk tujuan ini	Penelitian ini menyoroti bahwa meskipun berbagai penelitian telah menerapkan NER dalam Hadis, hanya sedikit yang mengoptimalkan BERT untuk identifikasi perawi. Model lain, seperti SVM dan Naive Bayes, telah digunakan dalam konteks yang berbeda, makalah ini sama sama menngunakan BERT sebagai topik modeling merubah kata kedalam bentuk vector
4	Rethinking of BERT Sentence Embedding for Text Classification	Galal, O. (2024). https://doi.org/10.21203/rs.3.rs-3920665/v1	Tujuan utama adalah untuk mengeksplorasi berbagai arsitektur agregasi untuk memanfaatkan penyematan lapisan akhir dan hidden layer embeddings dalam tugas klasifikasi teks, khususnya analisis sentimen dan deteksi sarkasme	Penelitian ini menyimpulkan bahwa freezing BERT dapat mengungguli fine-tuning BERT untuk tugas klasifikasi teks. Penelitian ini juga menyoroti efektivitas pembelajaran multitugas dalam meningkatkan kinerja pada tugas-tugas yang menantang seperti deteksi sarkasme	Penulis menyarankan agar komunitas penelitian berfokus pada penggunaan model yang telah dilatih sebelumnya seperti BERT sebagai ekstraktor fitur daripada menyempurnakannya, karena pendekatan ini dapat menghasilkan kinerja yang lebih baik dan mengurangi waktu pelatihan	Penelitian ini berfokus pada analisis sentimen dan deteksi sarkasme, penerapan BERT dalam klastering Hadis juga dapat memperoleh manfaat dari arsitektur agregasi yang diusulkan dan strategi freezing BERT, yang berpotensi menghasilkan akurasi klastering yang lebih baik dalam teks-teks keagamaan.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Devlin, J. (2018). arXiv. https://doi.org/10.48550/arxiv.1810.04805	untuk memperkenalkan model representasi bahasa baru yang melatih representasi dwiarah mendalam dari teks yang tidak berlabel, yang memungkinkannya untuk secara bersamaan mengondisikan pada konteks kiri dan kanan di semua lapisan	BERT mencapai hasil mutakhir pada sebelas tugas NLP, yang secara signifikan meningkatkan metrik kinerja seperti skor GLUE dan tolok ukur jawaban pertanyaan SquAD	Penelitian ini menyarankan bahwa dwiarah dan dua tugas prapelatihan sangat penting untuk peningkatan kinerja. Namun, makalah ini mencatat bahwa teknik saat ini dapat membatasi potensi representasi pra-pelatihan, khususnya dalam pendekatan penyempurnaan	Meskipun penelitian ini tidak secara khusus membahas klastering Hadis, kemampuan BERT untuk menangani berbagai tugas NLP, termasuk menjawab pertanyaan, menunjukkan bahwa BERT dapat diterapkan secara efektif untuk mengklasifikasikan teks Hadis dari sumber-sumber seperti Sahih Bukhari dan Muslim.
6.	An Unsupervised Sentence Embedding Method by Mutual Information Maximization	Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, Lidong Bing. arXiv. 2020 https://doi.org/10.48550/arXiv.2009.12061	Tujuan utama adalah untuk mengusulkan model penyisipan kalimat tanpa pengawasan yang baru, bernama Info-Sentence BERT (IS-BERT)	IS-BERT secara signifikan mengungguli baseline tanpa pengawasan lainnya pada tugas Semantic Textual Similarity (STS) dan SentEval, dan kompetitif dengan metode yang diawasi dalam skenario di mana data berlabel langka	Kelemahan yang perlu diperhatikan adalah bahwa meskipun IS-BERT berkinerja baik dalam pengaturan tanpa pengawasan, IS-BERT mungkin tidak selalu mengungguli SBERT saat disempurnakan pada data berlabel	Menggunakan metode IS-BERT yang dapat menawarkan keuntungan dalam skenario di mana jika menggunakan data Tanpa label.
7.	Fabricated Hadith Detection:	K. Gaanoun dan M. Alsuhaibani, 2022. IEEE Access , vol.	Penelitian ini bertujuan untuk mendeteksi hadis-	Sistem yang diusulkan berdasarkan model bahasa transformator	Keunggulannya dapat mendeteksi derajat dan jenis hadis, seperti	Penelitian ini dilskuksn untuk pengklasifikasian pembelajaran mesin untuk

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	A Novel Matn-Based Approach With Transformer Language Models	10, hlm. 113330-113342, 2022, doi: 10.1109/ACCESS.2022.3217457	hadis palsu (Mawdu) dengan berfokus pada teks sentral (Matn) dan bukan pada rantai perawi (Sanad)	(BERT) mencapai skor MH F1 terkini sebesar 92,47% Penelitian ini menyoroti keunggulan model bahasa transformator (TLM) Arab dibandingkan model pembelajaran mesin klasik untuk autentikasi hadis	mengklasifikasikan hadis Daif ke dalam beberapa jenis. Dan menyarankan pengembangan model yang lebih terspesialisasi yang dilatih pada korpus hadis untuk sistem analisis yang lebih baik	klasifikasi hadis, tetapi penelitian ini secara unik berfokus pada Matn menggunakan TLM, yang menandai kemajuan signifikan di bidang tersebut.
8.	Classification Performance Comparison of BERT and IndoBERT on Self-Report of COVID-19 Status on Social Media	Budiman, I., Faisal, M. R., Faridhah, A., Farmadi, A., Mazdadi, M. I., Saragih, T. H., ... & Abadi, F. (2024). Journal of Computer Sciences Institute, 30, 61-67. https://doi.org/10.35784/jcsi.5564	Tujuan utama adalah untuk membandingkan efektivitas model BERT dan IndoBERT dalam mengidentifikasi pesan laporan mandiri status COVID-19 dari media sosial, menggunakan data teks mentah dan yang telah diproses sebelumnya	Model IndoBERT mengungguli model BERT, mencapai akurasi 94%, sedangkan BERT mencapai akurasi 82%. Model IndoBERT juga memiliki spesifisitas yang lebih tinggi (92%) dan sensitivitas (96%) dibandingkan dengan spesifisitas BERT (89,11%) dan sensitivitas (74,75%)	Penelitian ini menyarankan eksplorasi lebih lanjut terhadap metodologi turunan BERT tambahan untuk meningkatkan kinerja klasifikasi. Kelemahannya dampak praproses teks pada kinerja model ditemukan tidak signifikan	Meskipun penelitian ini berfokus pada klasifikasi status COVID-19, BERT juga bisa diterapkan di domain lain, seperti klasterisai Hadis. Efektivitas BERT dalam berbagai tugas NLP, termasuk klasterisasi, memungkinkan untuk menjalankan dalam berbagai tugas, sehingga bisa diterapkan dalam penelitian ini
9.	Classification of Hadith	HM Abdelaal, BR Elemary dan HA Youness, IEEE	Tujuan utama dari penelitian ini adalah untuk	Penelitian ini bertujuan untuk membangun model pengklasifikasi	Penelitian ini menyimpulkan bahwa pengklasifikasi Decision	Fokus penelitian ini berfokus pada algoritma pembelajaran terbimbing

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	According to Its Content Based on Supervised Learning Algorithms	Access , vol. 7, hlm. 152379-152387, 2019, doi: 10.1109/ACCESS.2019.2948159.	mengklasifikasikan hadis secara otomatis ke dalam beberapa kategori berdasarkan isinya, khususnya berfokus pada teks hadis sambil mengabaikan Sanad (perawi) karena dianggap tidak relevan untuk tujuan klasifikasi	yang dapat membedakan antara kategori seperti Tauhid, Puasa, Shalat, Zakat, dan Haji menggunakan teknik penambangan data dan pembelajaran mesin	Tree (DT) mencapai akurasi tertinggi di antara pengklasifikasi yang diuji,. Makalah ini menyarankan perlunya penelitian lebih lanjut untuk meningkatkan proses klasifikasi, mungkin dengan mengeksplorasi algoritma pembelajaran mesin tambahan atau pendekatan hibrida.	tradisional, pengklasifikasi Decision Tree (DT) yang berpotensi yang lebih unggul dalam pengelompokan hadis dibanding metode yang dibahas dalam penelitian ini. Perbedaannya terletak pada metode yang dipakai yaitu supervised. Dan persamaannya adalah sama sama dalam penglompokan hadis.
10.	Classification of Bulughul Maraam Categories: Prohibitions, Recommendations, and Information Using Extreme Learning Machine and Fasttext	Handayani ,Najiyah ,Wisnuwardana Tahun 2023, Jurnal Online Informatika, DOI: https://doi.org/10.15575/join.v8i2.1205	Tujuan utama penelitian ini adalah mengklasifikasikan hadis dari kitab Bulughul Maraam ke dalam kategori larangan, anjuran, dan informasi.	Penelitian ini berhasil mengklasifikasikan hadis dengan akurasi 86,31%, yang merupakan peningkatan dari penelitian sebelumnya yang menggunakan metode lain seperti RCNN dan Naive Bayes, yang mencapai akurasi lebih rendah	Penulis menyarankan agar penelitian lebih lanjut dapat mengeksplorasi algoritma tambahan dan teknik praproses data untuk lebih meningkatkan akurasi klasifikasi.	Penelitiann ini sama sama meneliti untuk mengelompokan Hadis, namun hadis ini dikhususkan pada hadis di kitab Bulughul Maraam, sedangkan perbedaanya pada metode dan sumber hadis yang akan diteliti.
11	The performance of BERT as data	Subakti, A., Murfi, H. & Hariadi, N. J Big Data 9, 15	Penelitian ini bertujuan untuk mengevaluasi	Temuan menunjukkan bahwa BERT secara signifikan mengungguli	Penelitian ini menyarankan eksplorasi lebih lanjut tentang aplikasi BERT	Penelitian ini membandingkan kinerja BERT dengan metode

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	representation of text clustering	(2022). https://doi.org/10.1186/s40537-022-00564-9	efektivitas BERT sebagai metode representasi teks untuk tugas klasterisasi, khususnya dalam skenario pembelajaran tanpa pengawasan.	TFIDF dalam merepresentasikan data teks untuk klasterisasi, mencapai hasil yang lebih baik dalam 28 dari 36 metrik yang dievaluasi. Hal ini menunjukkan bahwa BERT secara efektif memposisikan teks yang serupa lebih dekat satu sama lain.	dalam pembelajaran tanpa pengawasan, karena sebagian besar penelitian telah difokuskan pada tugas yang diawasi. Kelemahan yang perlu diperhatikan adalah variabilitas kinerja berdasarkan metode ekstraksi fitur dan normalisasi yang berbeda, yang dapat memengaruhi hasil klasterisasi	tradisional seperti TF-IDF dalam aplikasi klasterisasi teks.
12	Topic modeling for scientific articles: exploring optimal hyperparameter tuning in bert	Wijanto, M. (2024). International Journal on Advanced Science Engineering and Information Technology, 14(3), 912-919. https://doi.org/10.18517/ijaseit.14.3.19347	Penelitian ini bertujuan untuk mengeksplorasi teknik pemodelan topik tingkat lanjut, khususnya pendekatan berbasis BERT, untuk meningkatkan analisis artikel ilmiah.	Penelitian ini menyimpulkan bahwa menggabungkan RoBERTa untuk penyisipan kata, PCA untuk reduksi dimensi, dan K-Means untuk klasterisasi menghasilkan hasil terbaik dalam hal nilai koherensi dan waktu eksekusi	Penelitian ini menyarankan bahwa metrik evaluasi lebih lanjut diperlukan untuk pemodelan topik berbasis BERT, karena evaluasi saat ini terbatas dibandingkan dengan metode tradisional seperti LDA	Penelitian ini berfokus pada pencarian parameter terbaik dalam menggunakan BERt sebagai metode. Schingga dengan kesimpulan bahwa K-Means dalam klasterisasi lebih unggul maka untuk klasterisasi bisa diterapkan pada penelitian klasterisasi Hadis Bukhari Muslim.
13	An integrated clustering and BERT	George, L., Sumathy, P. An integrated clustering and BERT	Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi	Penelitian ini berhasil mengembangkan dan mengevaluasi kerangka kerja pemodelan topik	Salah satu kelemahannya adalah kompleksitas komputasi yang melekat pada algoritma clustering	Penelitian ini fokus pada mengembangkan dan mengevaluasi kerangka kerja pemodelan topik

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	framework for improved topic modeling	framework for improved topic modeling. Int. j. inf. tecnol. 15, 2187–2195 (2023). https://doi.org/10.1007/s41870-023-01268-w	kerangka kerja pemodelan topik yang ditingkatkan, dengan fokus pada kombinasi teknik elustering danBERT (Representasi Encoder Bidirectional dari Transformers).	terintegrasi yang menggabung kanclustering dan model BERT-LDA (Representasi Encoder Bidirectional dari Transformers - Latent Dirichlet Allocation)	ketiku dihadapkan dengan jumlah fitur yang tinggi, meskipun pengurangan dimensi seperti PCA, t-SNE, dan UMAP digunakan untuk mitigasi. Sebagai saran untuk pekerjaan di masa depan, diharap meningkatkan eksplorasi lebih lanjut terhadap varian BERT yang lebih canggih.	terintegrasi yang menggabung kanclustering dan modelBERT-LDA. Sedangkan pada klasterisasi bukhari muslim mengembangkan metode BERT dengan TF-IDF
14	A Hybrid Model Integrating LDA, BERT, and Clustering for Enhanced Topic Modeling	Yadav, A.K., Gupta, T., Kumar, M. et al. A Hybrid Model Integrating LDA, BERT, and Clustering for Enhanced Topic Modeling. (2025) https://doi.org/10.1007/s11135-025-02077-y	Tujuan utama dari penelitian ini adalah untuk mengevaluasi berbagai metode penambangan teks untuk mengidentifikasi topik dalam korpus teks dengan mengungkap struktur semantik tersembunyi.	Model ini menggabungkan kekuatan Latent Dirichlet Allocation (LDA) dengan representasi kontekstual mendalam dari Representasi Encoder Bidirectional dari Transformers (BERT), serta mengintegrasikan teknikclustering dan pengurangan dimensi.	Sangat sensitif terhadap penyesuaian hyper-parameter yang ekstensif. Untuk meningkatkan kinerja lebih baik disarankan untuk menggabungkan lebih banyak embedding deep learning, eksplorasi model transformer canggih dan algoritma deep learning lainnya.	Pada penelitian ini fokus pada penggabungan dua metode yaitu BERT dan LDA sebagai metode terbaik yang dihasilkan, sedangkan dalam klasterisasi Hadis Bukhari Muslim fokus pada penerapan metode BERT - TF-IDF pada Hadis Bukhari Muslim.

2.1. Landasan Teori

2.3.1 Model BERT

Model *Bidirectional Encoder Representations from Transformers* (BERT) merupakan kemajuan signifikan dalam bidang pemrosesan bahasa alami (NLP). Dikembangkan oleh para peneliti di Google, BERT dibedakan oleh kemampuannya untuk menghasilkan representasi dua arah yang mendalam, yang memungkinkannya untuk mempertimbangkan konteks dari sisi kiri dan kanan sebuah kata secara bersamaan. Ini kontras dengan model sebelumnya yang memproses teks secara searah, yang menyebabkan keterbatasan dalam memahami konstruksi bahasa yang bernuansa. Arsitektur BERT didasarkan pada model Transformer, yang menggunakan mekanisme perhatian multi-kepala yang memungkinkan model untuk fokus pada bagian yang berbeda dari urutan input, dengan demikian menangkap hubungan yang kompleks dalam data [35] [36] .

Fase pra-pelatihan BERT melibatkan pelatihan pada sejumlah besar teks yang tidak berlabel, yang memungkinkannya mempelajari representasi bahasa yang kaya. Ini dicapai melalui dua tugas utama: pemodelan bahasa bertopeng dan prediksi kalimat berikutnya. Dalam pemodelan bahasa bertopeng, kata-kata acak dalam sebuah kalimat ditutupi, dan model belajar untuk memprediksi kata-kata bertopeng ini berdasarkan konteksnya. Prediksi kalimat berikutnya melibatkan penentuan apakah kalimat yang diberikan secara logis mengikuti yang lain, yang membantu model memahami hubungan kalimat [35][37] . Pendekatan pelatihan ganda ini membekali BERT dengan pemahaman bahasa yang kuat, membuatnya

sangat efektif untuk berbagai tugas hilir, termasuk klasifikasi teks, menjawab pertanyaan, dan pengenalan entitas bernama [37].

Fleksibilitas BERT lebih ditingkatkan oleh kemampuan fine-tuning-nya. Setelah fase pra-pelatihan, BERT dapat disempurnakan pada tugas-tugas tertentu dengan modifikasi arsitektur tambahan yang minimal. Kemampuan beradaptasi ini telah menghasilkan BERT yang mencapai hasil mutakhir di berbagai tolok ukur NLP, termasuk tolok ukur General Language Understanding Evaluation (GLUE), yang mengungguli model sebelumnya secara signifikan [35]. Kemampuan untuk menyempurnakan BERT untuk aplikasi tertentu telah menjadikannya model dasar dalam lanskap NLP, yang menginspirasi berbagai variasi dan adaptasi, seperti RoBERTa dan DistilBERT, yang bertujuan untuk meningkatkan efisiensi dan kinerja [38].

Arsitektur BERT dicirikan oleh penggunaan beberapa lapisan blok Transformer, yang masing-masing terdiri dari mekanisme self-attention dan jaringan saraf feed-forward. Pendekatan berlapis ini memungkinkan BERT untuk membangun representasi teks input yang semakin abstrak, yang menangkap informasi sintaksis dan semantik [35] [39]. Mekanisme perhatian khususnya perlu diperhatikan, karena memungkinkan model untuk mempertimbangkan pentingnya kata-kata yang berbeda dalam sebuah kalimat secara dinamis, memfasilitasi pemahaman konteks yang lebih bernuansa [40]. Kemampuan ini sangat penting untuk tugas-tugas yang memerlukan pemahaman bahasa yang mendalam, seperti deteksi metafora dan analisis sentimen, di mana maknanya dapat berubah secara dramatis berdasarkan konteks[41].

Selain itu, dampak BERT melampaui pemrosesan bahasa Inggris. Varian multibahasa BERT, yang dikenal sebagai mBERT, telah menunjukkan kemampuan lintas bahasa yang mengesankan, yang memungkinkannya untuk bekerja dengan baik pada tugas-tugas dalam berbagai bahasa tanpa perlu pelatihan ulang yang ekstensif [42]. Fitur ini khususnya bermanfaat dalam dunia yang mengglobal di mana keragaman bahasa merupakan faktor penting dalam komunikasi dan penyebaran informasi. Kemampuan BERT untuk melakukan generalisasi lintas bahasa menggarisbawahi peran mendasarnya dalam pengembangan model NLP masa depan [43][42] .

Penerapan BERT dalam berbagai domain juga patut diperhatikan. Misalnya, dalam bidang genomik, adaptasi BERT, seperti DNABERT, telah dikembangkan untuk menganalisis sekuens DNA, yang menggambarkan fleksibilitas model di luar pemrosesan teks tradisional [44]. Demikian pula, BERT telah digunakan dalam analisis media sosial untuk mendeteksi misinformasi dan teori konspirasi, yang menunjukkan relevansinya dalam masalah masyarakat kontemporer [45][46]. Aplikasi ini menyoroti potensi model untuk mengatasi tantangan kompleks di berbagai bidang, mulai dari perawatan kesehatan hingga ilmu sosial.

Terlepas dari keberhasilannya, BERT bukannya tanpa keterbatasan. Ukuran model yang besar dan persyaratan komputasi dapat menimbulkan tantangan untuk penerapan di lingkungan dengan keterbatasan sumber daya. Selain itu, interpretabilitas prediksi BERT masih menjadi topik penelitian yang sedang berlangsung, karena memahami proses pengambilan keputusan model pembelajaran mendalam sangat penting untuk kepercayaan dan akuntabilitas dalam

aplikasi AI [43] [40]. Para peneliti secara aktif mengeksplorasi metode untuk meningkatkan interpretabilitas BERT dan model serupa, yang bertujuan untuk menjembatani kesenjangan antara kinerja dan pemahaman [43].

Sebagai kesimpulan, model BERT telah mengubah lanskap pemrosesan bahasa alami secara mendasar melalui arsitektur dan metodologi pelatihannya yang inovatif. Kemampuannya untuk menghasilkan representasi yang sadar konteks, ditambah dengan kemampuan adaptasinya terhadap berbagai tugas, telah menjadikannya sebagai landasan penelitian dan aplikasi NLP modern. Seiring dengan terus berkembangnya bidang ini, pengaruh BERT kemungkinan akan terus berlanjut, menginspirasi model dan pendekatan baru yang dibangun di atas prinsip-prinsip dasarnya.

2.3.2. Natural Language Processing NLP

Natural Language Processing (NLP) adalah bidang multifaset yang bersinggungan dengan berbagai domain, termasuk kecerdasan buatan (AI), linguistik, dan ilmu komputer. Ini mencakup metodologi dan teknologi yang memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa manusia dengan cara yang bermakna. Evolusi NLP telah ditandai oleh kemajuan signifikan dalam teknik komputasi, yang telah mengubahnya dari pengejaran teoritis menjadi alat praktis dengan aplikasi dunia nyata yang luas. Sintesis ini akan mengeksplorasi dasar-dasar teoritis NLP, konteks historisnya, kemajuan saat ini, dan arah masa depan, yang didukung oleh literatur yang relevan.

Prinsip dasar NLP berakar dalam pada studi linguistik dan ilmu kognitif. NLP melibatkan otomatisasi analisis bahasa, yang meliputi sintaksis, semantik, dan pragmatik. Menurut Tilton dan Arnold, NLP pada dasarnya adalah tentang interaksi antara dialek manusia dan sistem komputer, yang memerlukan pemahaman yang komprehensif tentang bahasa kognitif dan padanan komputasinya [47]. Hirschberg dan Manning lebih lanjut menjelaskan bahwa pendekatan komputasi awal difokuskan pada otomatisasi analisis struktur linguistik, yang mengarah pada pengembangan teknologi penting seperti terjemahan mesin dan pengenalan ucapan [48] . Teknologi dasar ini meletakkan dasar untuk aplikasi yang lebih canggih, termasuk sistem dialog lisan dan alat analisis sentimen, yang sekarang lazim di berbagai sektor.

Hubungan antara NLP dan AI sangat signifikan, karena kemajuan dalam pembelajaran mesin dan pembelajaran mendalam telah mendorong kemampuan sistem NLP. Keezhatta menekankan bahwa integrasi aplikasi AI ke dalam platform NLP telah meningkatkan efektivitasnya dalam melakukan tugas linguistik yang kompleks, seperti penguraian dan analisis semantik [49]. Sinergi ini telah menghasilkan munculnya model NLP yang kuat yang dapat belajar dari kumpulan data yang luas, sehingga meningkatkan akurasi dan efisiensinya dalam memahami bahasa manusia. Transisi dari kerangka teoritis ke aplikasi praktis terbukti dalam karya Jin dkk., yang menyoroti dampak sosial dari teknologi NLP dan mengadvokasi pengembangannya dengan fokus pada kebaikan sosial [50].

Seiring terus berkembangnya NLP, metodologi yang digunakan di bidang ini juga semakin beragam. Studi terkini telah mengeksplorasi penggunaan

pembelajaran transfer, yang telah merevolusi pelatihan model NLP. Jiang et al. membahas bagaimana penyempurnaan model yang telah dilatih sebelumnya pada tugas-tugas tertentu dapat meningkatkan kinerjanya sekaligus mengatasi tantangan *overfitting* dan *generalisasi* [51]. Pendekatan ini telah menjadi landasan NLP modern, yang memungkinkan para peneliti untuk memanfaatkan kumpulan data skala besar untuk berbagai aplikasi, mulai dari perawatan kesehatan hingga analisis media sosial.

Kerangka kerja teoritis yang memandu penelitian NLP juga berkembang. Program *Minimalis Chomsky* menawarkan perspektif linguistik yang dapat menginformasikan metodologi NLP, khususnya dalam bidang sintaksis dan tata bahasa [52]. Dengan menerapkan prinsip-prinsip dari tata bahasa generatif, para peneliti dapat mengembangkan model yang lebih canggih yang lebih baik dalam menangkap kompleksitas bahasa manusia. Landasan teoritis ini penting untuk memajukan teknologi NLP dan memastikan keselarasannya dengan realitas linguistik.

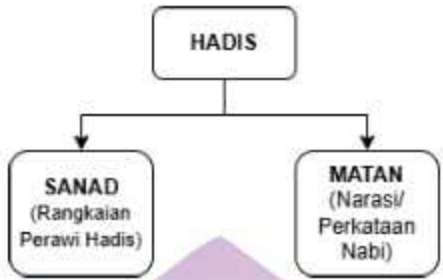
Sebagai kesimpulan, bidang Pemrosesan Bahasa Alami dicirikan oleh interaksi dinamis antara teori dan aplikasi. Integrasi AI, kemajuan dalam pembelajaran mesin, dan eksplorasi pertimbangan etika membentuk masa depan NLP. Ketika para peneliti terus menyempurnakan metodologi dan mengembangkan aplikasi inovatif, potensi NLP untuk mengubah berbagai sektor, dari perawatan kesehatan hingga pendidikan, tetap besar. Dialog yang sedang berlangsung antara kerangka kerja teoritis dan implementasi praktis akan sangat penting dalam

menavigasi tantangan dan peluang yang ada di depan dalam bidang yang berkembang pesat ini.

2.3.3 Hadis Sahih Bukhari dan Muslim

Dasar teori kumpulan hadis Sahih Bukhari dan Sahih Muslim berakar pada status keduanya sebagai sumber ajaran Islam paling autentik setelah Al-Qur'an. Kumpulan hadis ini sering disebut sebagai "*al-Sahihayn*," yang berarti "dua yang autentik," dan dikenal karena metodologinya yang ketat dalam penyusunan dan autentikasi hadis. Signifikansi teks-teks ini tidak hanya terletak pada isinya, tetapi juga pada proses ilmiah yang mendasari penerimaan mereka dalam tradisi Islam. Para ulama telah menetapkan bahwa hadis berfungsi sebagai sumber penting untuk memahami hukum Islam (Syariah) dan perilaku etis, melengkapi ajaran yang ditemukan dalam Al-Qur'an [53][54] [55].

Metodologi yang digunakan oleh Imam al-Bukhari dan Imam Muslim dalam menyusun kumpulan hadis mereka dicirikan oleh kriteria keaslian yang ketat. Ini termasuk evaluasi rantai perawi (Isnad) dan isi (Matan) setiap hadis seperti pada Gambar 2.1. mengenai Komponen Rangkaian Hadis. Sebuah hadis dianggap "Sahih" atau otentik jika memenuhi lima syarat penting, termasuk integritas dan keandalan perawi, serta kesinambungan rantai transmisi [56] [57]. Sifat teliti dari proses ini telah menyebabkan penerimaan luas terhadap koleksi ini di kalangan Muslim Sunni, yang menganggapnya penting untuk mendapatkan putusan hukum dan pedoman etika [58] [59].



Gambar 2.1. Komponen Rangkaian Hadis

Selain itu, peran hadis dalam yurisprudensi Islam tidak dapat dilebih-lebihkan. Ia berfungsi sebagai sumber hukum sekunder, yang memberikan konteks dan penjabaran pada ayat-ayat Al-Qur'an. Literatur hadis, khususnya yang terdapat dalam Sahih Bukhari dan Sahih Muslim, mengandung banyak ajaran yang membahas berbagai aspek kehidupan, termasuk moralitas, keadilan sosial, dan perilaku pribadi [60][61] . Misalnya, hadis menekankan pentingnya kesederhanaan, rasa hormat, dan kasih sayang, yang merupakan bagian integral dari kerangka etika Islam [62][63]. Ajaran yang berasal dari teks-teks ini bukan sekadar catatan sejarah, tetapi diterapkan secara aktif dalam kajian dan praktik Islam kontemporer.

Dasar teoritis Sahih Bukhari dan Sahih Muslim tertanam dalam pada keasliannya, ketelitian metodologis, dan relevansinya dengan hukum dan etika Islam. Koleksi-koleksi ini tidak hanya berfungsi sebagai sumber penting pengetahuan agama tetapi juga memainkan peran penting dalam membentuk tatanan moral dan sosial masyarakat Muslim. Studi dan interpretasi yang sedang berlangsung dari teks-teks ini terus menjadi bidang yang dinamis, yang mencerminkan sifat beasiswa Islam yang terus berkembang dan penerapannya dalam konteks kontemporer [7][53][64].

2.3.4 Big Data Hadis Sahih Bukhari dan Muslim

Integrasi teknologi big data ke dalam studi Hadis merupakan kemajuan signifikan di bidang studi Islam, khususnya dalam meningkatkan metodologi penelitian dan meningkatkan aksesibilitas literatur Hadis. Pengembangan kumpulan data yang dirancang khusus untuk disambiguasi perawi Hadis, seperti kumpulan data AR-Sanad 280K, mencontohkan bagaimana kecerdasan buatan dan pembelajaran mesin dapat dimanfaatkan untuk mengatasi tantangan kompleks dalam studi Hadis. Dataset ini, yang mencakup sejumlah besar label kelas, memungkinkan peneliti untuk menerapkan teknik pembelajaran mesin tingkat lanjut, termasuk menyempurnakan model berbasis BERT, untuk mencapai tingkat akurasi yang tinggi dalam mengklasifikasikan perawi [65].

Persinggungan antara big data dan kajian hadis tidak terbatas pada kemajuan teknis; hal itu juga mencakup dimensi etika dan filosofis dari penelitian hadis. Para cendekiawan semakin menyadari implikasi penggunaan teknologi dalam kajian agama, khususnya yang menyangkut keaslian dan penafsiran hadis. Wacana yang sedang berlangsung seputar kritik keaslian hadis oleh para pemikir Islam kontemporer mencerminkan semakin meningkatnya pengakuan akan perlunya metodologi yang ketat yang selaras dengan kajian tradisional dan kemampuan teknologi modern ([66]).

Selain itu, eksplorasi hadis dalam konteks isu-isu kontemporer, seperti tantangan media sosial yang dihadapi oleh Generasi Z, menggambarkan relevansi kajian hadis dalam menangani masalah-masalah masyarakat modern. Penerapan big

data dalam kajian hadis juga meluas ke ranah praktik pendidikan. Integrasi teknik pembelajaran mesin dalam penilaian hasil pendidikan dalam studi Al-Qur'an dan hadis telah menunjukkan hasil yang menjanjikan, dengan metodologi seperti penilaian autentik yang digunakan untuk mengevaluasi pembelajaran siswa secara efektif. Pendekatan ini tidak hanya meningkatkan pengalaman pendidikan tetapi juga sejalan dengan prinsip-prinsip pedagogi Islam, yang menekankan pentingnya perolehan pengetahuan dan pengembangan moral [67].

Lebih jauh lagi, penggunaan metodologi data besar dalam menganalisis implikasi sosial-politik hadis, khususnya yang menyangkut pemberdayaan perempuan dan partisipasi politik, telah membuka jalan baru untuk penelitian. Dengan memeriksa persepsi berbagai ulama Islam tentang isu-isu ini, para peneliti dapat memperoleh pemahaman yang lebih dalam tentang peran yang dimainkan hadis dalam membentuk pemikiran dan praktik Islam kontemporer [68].

Sebagai kesimpulan, integrasi teknologi data besar ke dalam studi hadis merupakan perubahan transformatif yang meningkatkan metodologi penelitian, meningkatkan aksesibilitas, dan membahas isu-isu kontemporer dalam konteks Islam. Kemajuan yang terus berlanjut dalam pembelajaran mesin, pemrosesan bahasa alami, dan klasifikasi data siap untuk mendefinisikan ulang lanskap kajian hadis, menjadikannya lebih tangguh, relevan, dan responsif terhadap kebutuhan masyarakat modern. Seiring para peneliti terus mengeksplorasi potensi big data di bidang ini, masa depan kajian hadis tampak menjanjikan, dengan kemungkinan mengungkap wawasan baru dan menumbuhkan pemahaman yang lebih dalam tentang ajaran Islam.

2.3.5 Topik Klastering HDBSCAN

Dalam pengklasteran data teks hasil representasi embedding seperti pada teks hadis Bukhari-Muslim, pemilihan algoritma klasterisasi memegang peran penting dalam menghasilkan topik yang representatif dan bermakna. Algoritma K-Means, meskipun populer karena kesederhanaannya dan kecepatan proses, memiliki keterbatasan penting, yaitu keharusan menentukan jumlah klaster di awal, asumsi bentuk klaster yang bulat *spherical,* serta ketidakmampuannya menangani outlier atau dokumen yang tidak cocok dalam klaster manapun [69]. Sebaliknya, algoritma HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) menawarkan keunggulan dalam menentukan jumlah klaster secara otomatis, mendeteksi noise/outlier, dan menangani distribusi klaster yang tidak beraturan, menjadikannya lebih andal untuk data teks yang telah diubah menjadi embedding vektor berdimensi tinggi, seperti hasil dari model BERT [70]. Dokumentasi resmi HDBSCAN juga menunjukkan bahwa metode ini lebih stabil dan akurat pada data dunia nyata dibandingkan K-Means, terutama pada data high-dimensional yang kompleks seperti embedding hadis [70] Oleh karena itu, dalam penelitian ini, HDBSCAN lebih dipilih karena kemampuannya menghasilkan klaster topik yang lebih natural, bermakna secara semantik, dan tidak terlalu dipengaruhi parameter awal yang sensitif seperti pada K-Means

2.3.6 Feature Extraction TF-IDF

Dalam penelitian topik modeling terhadap kumpulan hadis Bukhari-Muslim, representasi teks menjadi vektor numerik merupakan tahap krusial yang memengaruhi efektivitas klasterisasi. Salah satu pendekatan klasik yang masih banyak digunakan adalah TF-IDF (*Term Frequency–Inverse Document Frequency*), yaitu metode pembobotan kata yang mengukur pentingnya suatu kata dalam sebuah dokumen relatif terhadap seluruh korpus. TF-IDF memberikan nilai tinggi untuk kata-kata yang sering muncul dalam satu dokumen namun jarang muncul di dokumen lain, sehingga efektif dalam menonjolkan istilah-istilah khas dari masing-masing hadis [71]. Dalam konteks klasterisasi, TF-IDF dapat digunakan sebagai fitur vektor yang merepresentasikan setiap dokumen, dan selanjutnya digunakan oleh algoritma clustering seperti K-Means atau HDBSCAN [72]. Meskipun TF-IDF tidak mempertimbangkan struktur semantik dan konteks makna seperti BERT, ia tetap relevan untuk ekstraksi fitur interpretable dan dapat digunakan secara gabungan dengan representasi embedding lainnya untuk memperkaya informasi, sebagaimana diterapkan pada pengklasteran hadis berbasis fitur campuran. Kombinasi embedding kontekstual dari BERT dengan bobot semantik dari TF-IDF terbukti dapat meningkatkan kualitas topik yang dihasilkan dan memperjelas pemisahan antar klaster [14].

2.3.7 Reduksi Dimensi PCA

Dalam penelitian pengklasteran teks berbasis embedding seperti pada kumpulan hadis Bukhari-Muslim, tantangan utama yang dihadapi adalah tingginya

dimensi representasi vektor yang dihasilkan oleh model bahasa seperti BERT. Dimensi tinggi ini dapat menyulitkan proses klasterisasi karena menyebabkan hilangnya makna jarak antar titik serta meningkatnya kompleksitas komputasi. Salah satu solusi yang umum digunakan untuk mengatasi hal ini adalah Principal Component Analysis (PCA), yaitu teknik reduksi dimensi linier yang memproyeksikan data ke ruang berdimensi lebih rendah dengan tetap mempertahankan sebanyak mungkin variansi aslinya [73]. PCA bekerja dengan mengidentifikasi komponen utama yang menjelaskan sebagian besar variasi dalam data, sehingga memungkinkan proses klasterisasi menjadi lebih efisien dan stabil. Meskipun PCA tidak mempertahankan struktur non-linier seperti halnya UMAP atau t-SNE, PCA tetap banyak digunakan karena sifatnya yang deterministik, cepat, dan mudah diinterpretasikan [74]. Dalam konteks penelitian ini, PCA digunakan sebagai tahap prapemrosesan sebelum diterapkannya algoritma klasterisasi seperti K-Means atau HDBSCAN, untuk menyederhanakan struktur data hasil embedding dan meningkatkan kualitas pemisahan klaster topik dalam kumpulan hadis.

2.3.8 Evaluasi Silhouette Score

Dalam evaluasi kualitas hasil klasterisasi terhadap data teks hasil representasi vektor, seperti hadis Bukhari-Muslim yang diubah menjadi embedding melalui model BERT, diperlukan metrik objektif untuk mengukur koherensi internal dan pemisahan antar-klaster. Salah satu metrik yang paling umum digunakan adalah *Silhouette Score*, yang mengukur seberapa mirip sebuah dokumen dengan dokumen-dokumen dalam klasternya sendiri dibandingkan

dengan dokumen di klaster lain [75]. Nilai Silhouette berkisar antara -1 hingga +1, di mana nilai lebih tinggi menunjukkan bahwa dokumen lebih cocok berada di klaster tersebut dan memiliki batas yang jelas dengan klaster lain. Dalam konteks penelitian topik modeling hadis, Silhouette Score dapat memberikan indikasi kuantitatif atas ketepatan model klasterisasi seperti K-Means atau HDBSCAN dalam memisahkan topik-topik teks yang secara semantik berbeda. Selain itu, metrik ini juga sangat berguna saat membandingkan performa antar metode klasterisasi pada hasil embedding teks, karena dapat menghitung kualitas klaster tanpa bergantung pada label ground truth, yang umumnya tidak tersedia dalam data hadis [76]. Oleh karena itu, *Silhouette Score* digunakan dalam penelitian ini sebagai alat validasi untuk memilih model klasterisasi yang paling representatif terhadap struktur topik dalam kumpulan hadis.

2.3.9 Evaluasi Davies-Bouldin Index (DBI)

Dalam proses evaluasi kualitas klasterisasi terhadap data teks hasil transformasi vektor seperti embedding hadis Bukhari-Muslim, penggunaan metrik internal sangat penting untuk menilai sejauh mana hasil klaster menggambarkan struktur semantik dokumen yang diolah. Salah satu metrik yang sering digunakan selain *Silhouette Score* adalah *Davies-Bouldin Index* (DBI). DBI mengukur rata-rata rasio antara jarak intra-klaster dan inter-klaster, di mana nilai yang lebih rendah menunjukkan klaster yang lebih kompak dan terpisah dengan baik [77]. Dengan kata lain, DBI mengasumsikan bahwa klaster yang baik memiliki variasi internal yang kecil dan jarak antar pusat klaster yang besar. Dalam konteks penelitian topik

modeling pada kumpulan hadis, DBI dapat digunakan untuk mengevaluasi seberapa baik algoritma seperti K-Means atau HDBSCAN berhasil membagi dokumen ke dalam kelompok topik yang saling berbeda secara semantik. DBI sangat bermanfaat ketika tidak tersedia label kategori topik sebagai ground truth, karena dapat memberikan penilaian kualitatif berbasis distribusi vektor embedding yang dihasilkan oleh model bahasa seperti BERT [78]. Oleh karena itu, dalam penelitian ini, *Davies-Bouldin Index* digunakan sebagai metrik pelengkap untuk mendukung analisis performa klasterisasi secara kuantitatif.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini termasuk ke dalam jenis penelitian empiris, mengacu pada pengumpulan data yang diperoleh dari penelitian-penelitian sebelumnya dan karakteristik dataset terdiri dari Shahih Bukhari dan Shahih Muslim memiliki sekitar 34.000 hadis shahih dengan pengulangan beberapa perawi yang berbeda yang tersimpan dalam file format csv.

Sifat dari penelitian ini sendiri adalah ekperimental. Dimana peneliti mencoba dengan model algoritma BERT untuk diterapkan pada klasterisasi topik hadis Bukhari Muslim dengan menggunakan beberapa metode .

Pendekatan kuantitatif digunakan pada penelitian ini, sesuai dengan yang akan diajukan untuk mengukur kinerja dari penerapan model algoritma BERT berdasarkan beberapa metode proses klasterisasi di dalamnya .

. Yang dilakukan secara sistematis mengikuti rencana yang terdefinisi dengan baik untuk mengumpulkan dan menganalisis data untuk memastikan validitas dan realibilitas hasil temuan. pemilihan metode penelitian yang tepat perlu diperhatikan beberapa karakteristik dataset sebelum masuk pada pemrosesan data.

3.2. Metode Pengumpulan Data

Sumber data yang dibutuhkan pada penelitian ini didapatkan dari metode *secondary data analysis* yang melibatkan pengguna data yang telah dikumpulkan

oleh peneliti-peneliti sebelumnya. Data yang didapatkan berupa arsip yang tersedia secara publik di Kaggle dan diperbolehkan untuk digunakan untuk penelitian berupa data format csv. Dari link https://www.kaggle.com/datasets/fahd09/hadith-dataset.

3.3. Metode Analisis Data

Untuk mencapai tujuan dan membuktikan hasil dari penelitian ini, maka beberapa tahapan metode analisis data dilakukan untuk mengetahui visualisasi dan akurasi pada penerapan model algoritma BERT untuk klasterisasi hadis dari Sahih Bukhari dan Muslim .

Penelitian ini menggunakan pendekatan pemodelan topik berbasis NLP (*Natural Language Processing*) yang mengintegrasikan representasi vektor embedding dan klasterisasi untuk menemukan struktur topik dalam kumpulan hadis Shahih Bukhari dan Muslim. Adapun tahapan metodologi dijelaskan secara rinci sebagai berikut:

Tahap pertama dalam penelitian ini adalah pengumpulan dan pemilihan dataset berupa kumpulan hadis dari kitab Shahih Bukhari dan Muslim. Dataset ini berisi teks hadis dalam bahasa Arab yang menjadi objek utama untuk proses analisis dan klasterisasi topik. Data ini disiapkan dalam format digital agar dapat diolah lebih lanjut melalui teknik NLP. Pemilihan hadis dari dua kitab utama ini bertujuan untuk menggali struktur tematik yang terkandung di dalamnya dan membentuk kelompok topik secara otomatis berdasarkan kemiripan semantik antar hadis.

Setelah data terkumpul, dilakukan tahap pra-pemrosesan untuk membersihkan dan menyiapkan teks hadis agar siap diolah secara komputasional. Tahapan ini mencakup penghapusan kata umum (stop word removal) yang tidak membawa makna penting, serta lemmatization untuk mengubah kata ke bentuk dasarnya. Proses ini penting agar hasil ekstraksi fitur dan pemodelan topik menjadi lebih akurat dan tidak bias oleh variasi bentuk kata.

Tahap selanjutnya adalah ekstraksi fitur dari teks hadis untuk membentuk representasi numerik. Dua teknik utama digunakan, yaitu TF-IDF (*Term Frequency Inverse Document Frequency*), yang menghitung bobot pentingnya kata dalam suatu dokumen terhadap keseluruhan korpus, serta word embedding berbasis BERT, yang mengubah teks hadis menjadi vektor berdimensi tinggi berdasarkan makna semantik kontekstual. Penggunaan kedua metode ini bertujuan untuk menangkap baik informasi frekuensi maupun kedalaman makna dari setiap teks.

Setelah representasi vektor diperoleh, dilakukan pemodelan topik untuk memetakan struktur semantik antar dokumen. Pemodelan ini mempertimbangkan panjang teks, nilai TF-IDF, dan embedding BERT, sehingga masing-masing hadis memiliki representasi vektor yang mencerminkan makna topik. Vektor-vektor ini digunakan sebagai dasar dalam proses klasterisasi. Sedangkan panjang teks dan TF-IDF harus di normalisasi terlebih dahulu sebelum digabung menjadi pemodelan topik gabungan.

Karena embedding BERT menghasilkan vektor berdimensi tinggi, dilakukan reduksi dimensi menggunakan PCA (*Principal Component Analysis*). Tujuan dari tahap ini adalah menyederhanakan kompleksitas data ke dalam dimensi

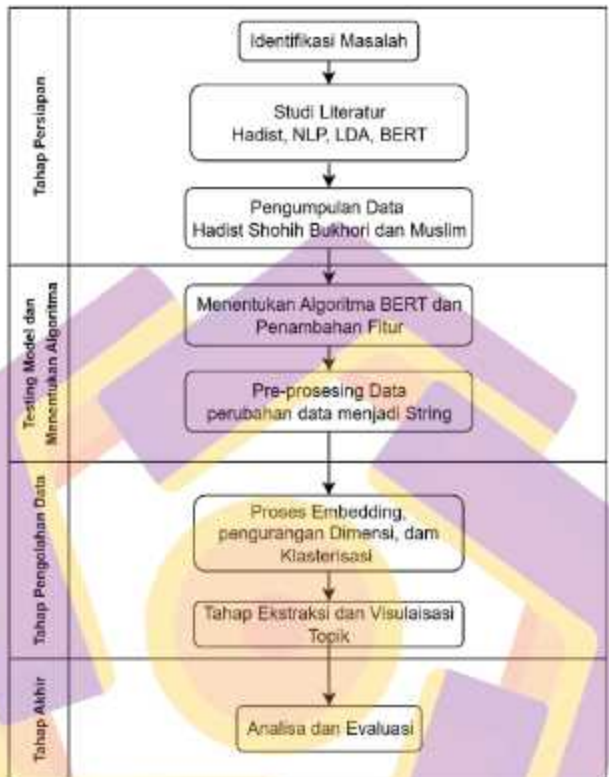
yang lebih rendah sambil mempertahankan informasi semantik utama. Hal ini juga meningkatkan efisiensi komputasi dan kemudahan visualisasi hasil klasterisasi.

Proses klasterisasi dilakukan menggunakan algoritma DBSCAN (*Density Based Spatial Clustering of Applications with Noise*). DBSCAN mampu membentuk klaster berdasarkan kepadatan data dan mendeteksi outlier yang tidak sesuai dengan kelompok manapun. Algoritma ini cocok untuk data hasil embedding karena tidak memerlukan penentuan jumlah klaster di awal dan mampu menangani klaster berbentuk arbitrer.

Tahap akhir dari metodologi ini adalah ekstraksi dan interpretasi topik dari setiap klaster yang terbentuk. Proses ini dilakukan dengan mengidentifikasi kata-kata kunci dominan dalam tiap kelompok untuk menyimpulkan tema utama yang mewakili klaster tersebut. Hasil akhir berupa daftar topik yang mencerminkan struktur tematik dalam kumpulan hadis Bukhari-Muslim, yang dapat digunakan untuk analisis lebih lanjut atau visualisasi pengetahuan keislaman.

3.4. Alur Penelitian

Untuk menyelesaiakn penelitian ini dibutuhkan langkah-langkah yang harus diselesaikan. Adapun langkah atau tahap yang dilakukan pada penelitian ini dapat dilihat pada keterangan dan Gambar 3.1 tentang alur penelitian.



Gambar 3.1. Alur Penelitian

1. Tahap Persiapan

   - Identifikasi Masalah

     Pada tahap awal, dilakukan perumusan masalah penelitian. Misalnya, bagaimana mengelompokkan topik-topik hadis dalam Shahih Bukhari dan Muslim secara otomatis menggunakan teknik NLP (Natural Language Processing).

Identifikasi ini penting untuk menentukan ruang lingkup penelitian, tujuan, serta manfaat yang akan diperoleh.

- Studi Literatur (Hadis, NLP, LDA, BERT)

  Mengkaji teori dan penelitian sebelumnya yang relevan, seperti:

  a. Konsep dasar hadis serta karakteristiknya.
  b. NLP untuk pengolahan bahasa alami.
  c. Algoritma topik modeling seperti LDA (*Latent Dirichlet Allocation*).
  d. Algoritma modern berbasis transformer (BERT) untuk representasi teks.
  e. Studi ini bertujuan memberikan dasar teori sekaligus menemukan gap penelitian.

- Pengumpulan Data (Hadis Shahih Bukhari dan Muslim)

  a. Mengambil dataset berupa teks hadis dari kitab Shahih Bukhari dan Muslim, baik dari sumber digital maupun manual.
  b. Data yang terkumpul akan menjadi bahan utama untuk proses pemodelan.

2. Tahap Testing Model dan Menentukan Algoritma

   - Menentukan Algoritma BERT dan Penambahan Fitur

     a. Menentukan model representasi teks yang digunakan, yaitu BERT (*Bidirectional Encoder Representations from Transformers*).

b. Menentukan fitur tambahan, seperti *stopword removal, tokenization, lemmatization, atau n-gram.*

c. Hal ini bertujuan agar data siap untuk dimasukkan ke dalam proses embedding.

- Pre-prosesing Data (Perubahan Data Menjadi String)

a. Melakukan pembersihan data, seperti: Menghapus tanda baca, angka, dan karakter yang tidak diperlukan.

b. Konversi teks menjadi format string yang seragam agar mudah diproses oleh BERT.

3. Tahap Pengolahan Data

- Proses Embedding, Pengurangan Dimensi, dan Klasterisasi

a. Embedding yaitu mengubah teks hadis menjadi representasi vektor numerik menggunakan model BERT.

b. Pengurangan dimensi Yaitu menggunakan metode seperti PCA atau UMAP untuk menyederhanakan dimensi vektor agar mudah divisualisasikan dan lebih efisien.

c. Klasterisasi yaitu mengelompokkan hadis berdasarkan kesamaan topik menggunakan algoritma clustering, misalnya K-Means atau HDBSCAN.

- Tahap Ekstraksi dan Visualisasi Topik

a. Melakukan ekstraksi kata-kata penting pada tiap klaster untuk mengetahui topik utama.

b. Visualisasi dilakukan menggunakan grafik, word cloud, atau peta topik sehingga hasil klasterisasi lebih mudah dipahami.

4. Tahap Akhir

- Analisa dan Evaluasi

a. Mengevaluasi hasil klasterisasi dengan mengukur kualitas cluster (menggunakan *silhouette score*, dan *Davies-Bouldin index*).

b. Melakukan analisis apakah topik yang dihasilkan relevan dengan isi hadis.

c. Memvisualisasikan hasil klasterisasi algoritma (BERT + HDBSCAN).

d. Menarik kesimpulan akhir terkait efektivitas metode yang digunakan serta memberikan rekomendasi untuk penelitian selanjutnya.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil Penelitaian

Penelitaian pengembangan model klusterisasi topik hadis Bukhari Muslim menggunakan BERT dengan penambahan fitur semantik dibahas melalui beberapa tahapan tahapan proses kerangka kerja seperti pada Gambar 4.1. di bawah ini.

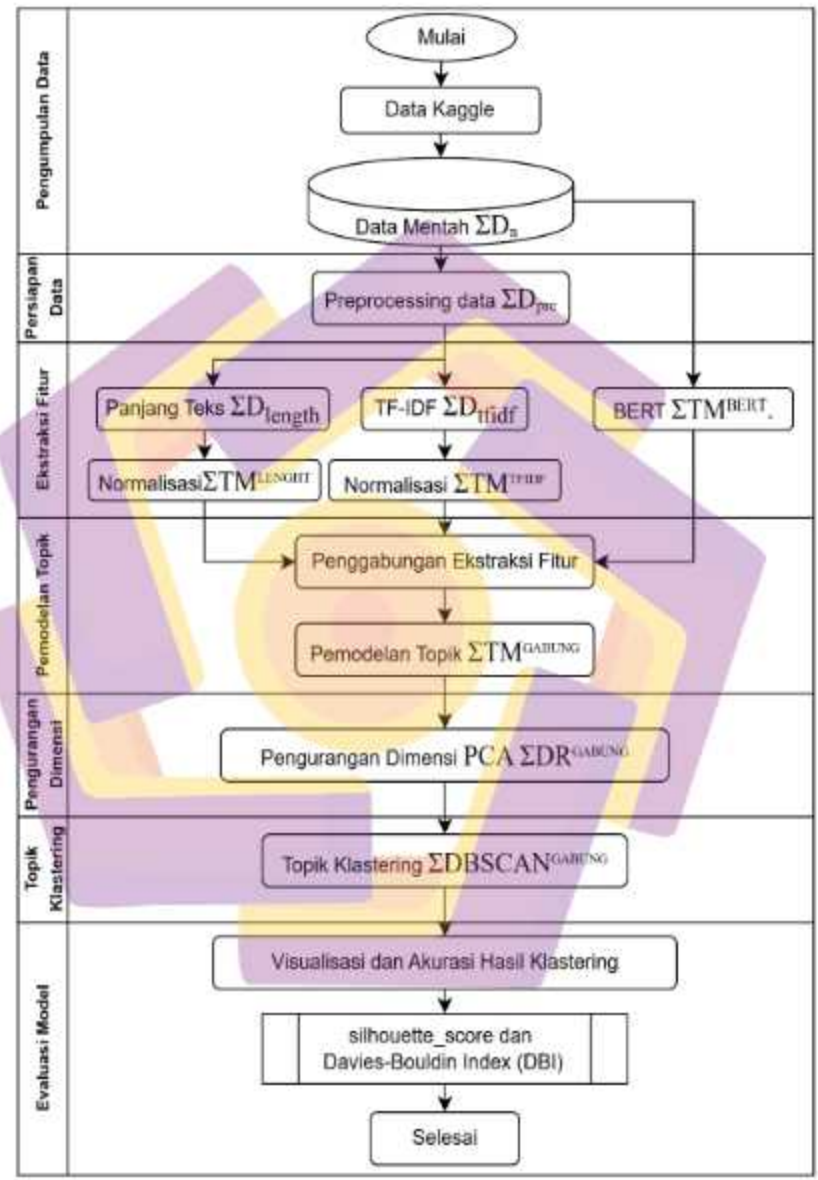
Gambar 4.1. Diagram Blok kerangka kerja pemodelan topik berbasis *Clustering* Integrasi Embedding BERT Dengan Fitur Semantik Tambahan Panjang Teks Dan TF-IDF

Diagram blok di atas pada Gambar 4.1 menggambarkan kerangka kerja pemodelan topik berbasis clustering yang menggabungkan Panjang Teks, TF-IDF,

dan BERT untuk mengelompokkan hadis dari kitab Shahih Bukhari dan Muslim. Proses dimulai dari input dataset, yaitu kumpulan teks hadis yang menjadi objek analisis. Tahap berikutnya adalah pra-pemrosesan, yang meliputi penghapusan stopword dan lemmatization untuk membersihkan serta menormalkan data teks agar siap diolah lebih lanjut. Setelah itu dilakukan feature extraction menggunakan metode Word Embedding dan TF-IDF untuk merepresentasikan teks dalam bentuk vektor numerik.

Selanjutnya, pada tahap pemodelan topik, digunakan kombinasi fitur Panjang Teks, TF-IDF, dan representasi vektor dari BERT untuk menangkap makna semantik dan konteks dari teks hadis. Agar hasil representasi lebih efisien, dilakukan pengurangan dimensi menggunakan PCA (Principal Component Analysis) guna mempertahankan informasi penting sambil mengurangi kompleksitas data. Tahap inti dari kerangka kerja ini adalah topik klastering menggunakan algoritma DBSCAN yang mampu mengidentifikasi kelompok topik berdasarkan kepadatan data tanpa perlu menentukan jumlah klaster di awal. Akhirnya, hasil klasterisasi divisualisasikan dan diukur tingkat akurasinya pada tahap visualisasi hasil dan akurasi, untuk memberikan Gambaran sebaran topik hadis serta kualitas pemodelan yang dihasilkan.

Sedangkan untuk proses ekseskusi program dalam mengeksekusi dataset dapat dilihat pada Gambar 4.2 mengenai diagram proses di bawah ini mulai dari awal sampai akhir. Dan adapun mengenai hasil masing masing langkah akan di jelaskan sesuai tahapan bagain bagian diagram proses pada Gambar 4.2.



Gambar 4.2. Diagram Proses

4.1.1 Pengumpulan Data

Pada tahap ini Gambar 4.2 Diagram Proses, data .csv dari kaggle menjadi sebuah dokumen input ΣD_n dari dataset hadis Bukhori Muslim dan dimana n merepresentasikan jumlah total dokumen hadis.

Dataset hadis dapat dilihat pada Gambar 4.3 yang menunjukkan bahwa terdapat kolom id hadis dan kolom isi dari hadis tersebut. Data Hadis bukhari muslim ini terdiri dari 34.441 hadis yang di masing masing kolom isi terdiri dari sanad dan matan di setiap hadisnya. Dengan panjang hadis bervariasi sesuai dengan jumlah kata pada sanat maupun jumlah kata pada sanadnya

id	hadith_id	text_ar
0	1	حدثنا الحميدي عبد الله بن الزبير ، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني محمد بن إبراهيم التيمي،
1	2	حدثنا عبد الله بن يوسف، قال أخبرنا مالك، عن هشام بن عروة، عن أبيه، عن عائشة أم المؤمنين رضي الله عنها أن الحارث
2	3	حدثنا يحيى بن بكير ، قال حدثنا الليث، عن عقيل، عن ابن شهاب، عن عروة بن الزبير، عن عائشة أم المؤمنين، أنها قالت أول
3	4	قال ابن شهاب وأخبرني أبو سلمة بن عبد الرحمن، أن جابر بن عبد الله الأنصاري، قال وهو يحدث عن فترة الوحي، فقال في
4	5	حدثنا موسى بن إسماعيل، قال حدثنا أبو عوانة، قال حدثنا موسى بن أبي عائشة، قال حدثنا سعيد بن جبير ، عن ابن عباس، في
5	6	حدثنا عياش، قال أخبرنا عبد الله، قال أخبرنا يونس، عن الزهري، ح وحدثنا بشر بن محمد، قال أخبرنا عبد الله، قال أخبرنا يو
6	7	حدثنا أبو اليمان الحكم بن نافع، قال أخبرنا شعيب، عن الزهري، قال أخبرني عبيد الله بن عبد الله بن عتبة بن مسعود أن عبد
0	8	حدثنا عبيد الله بن موسى، قال أخبرنا حنظلة بن أبي سفيان، عن عكرمة بن خالد، عن ابن عمر رضي الله عنهما قال قال رسو
1	9	حدثنا عبد الله بن محمد، قال حدثنا أبو عامر العقدي، قال حدثنا سليمان بن بلال، عن عبد الله بن دينار ، عن أبي صالح، عن أبي
2	10	حدثنا آدم بن أبي إياس، قال حدثنا شعبة، عن عبد الله بن أبي السفر ، وإسماعيل، عن الشعبي، عن عبد الله بن عمرو رضي الله

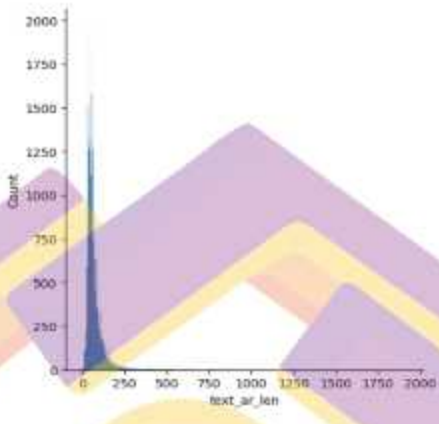
Gambar 4.3. Dataset Hadis Bukhori Muslim

4.1.2 Persiapan Data

Tahap ini menghasilkan dokumen yang telah dipra-proses ΣD_{pre} dari ΣD_n dengan melakukan langkah-langkah praproses.

Pada Gambar 4.4 di bawah ini Distribusi Grafik menunjukkan distribusi yang sangat condong ke kiri (*left-skewed*). Sebagian besar teks memiliki panjang pendek (sekitar 0–250) dan semakin panjang teks, semakin sedikit frekuensinya.

Sehingga data teks yang dianalisis sebagian besar pendek, hanya sedikit data yang panjang teksnya di atas 500.



Gambar 4.4. Distribusi panjang Teks Hadis

Pada pra-pemrosesan dataset merupakan tahap awal yang sangat penting dalam proses analisis teks hadis Bukhari-Muslim, karena menentukan kualitas data yang akan digunakan pada tahap pemodelan selanjutnya. Sehingga dipastikan tidak ada data duplikasi, data kosong, maupun tidak sama jenis datanya. Sehingga untuk memasitikan data disini sama jenisnya maka data dirubah ke dalam bentuk tipe string agar mudah untuk dioleh dalam pemrosesan algoritma BERT.

Disamping merubah data ke dalam bentuk string, pada bagian pra-pemrosesan ini. dilakukan dua proses lanjutan, yaitu Stop Word Removal dan Lemmatization.

Tahap Stop Word Removal merupakan bagian dari proses pra-pemrosesan teks yang bertujuan untuk menghapus kata-kata umum (stopwords) yang tidak memiliki makna signifikan terhadap konteks semantik suatu kalimat. Dalam

penelitian ini, tahap ini diterapkan pada teks hadis berbahasa Arab yang terdapat pada kolom text_Ar dalam dataset hadis Bukhari-Muslim.

Stopwords dalam bahasa Arab mencakup kata-kata seperti من (dari), في (di), على (atas), إلى (ke), عن (tentang), ما (apa), لا (tidak), إن (sesungguhnya), أن (bahwa), كان (adalah) dan kata umum lainnya yang sering muncul dalam kalimat namun tidak membawa informasi tematik penting. Kata-kata tersebut dihapus agar model tidak memberikan bobot semantik yang tidak relevan terhadap fitur teks yang digunakan untuk klasterisasi topik.

Dalam penelitian ini, proses penghapusan stopwords dilakukan dengan menggunakan daftar kata umum bahasa Arab yang diadaptasi dari Arabic Stopwords List (NLTK), kemudian dilakukan penyaringan (*filtering*) dengan menghapus kata-kata yang termasuk dalam daftar stopwords.

Sebagai contoh, kalimat asli dari hadis:

حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني " محمد بن إبراهيم التيمي، أنه سمع علقمة بن وقاص الليثي، يقول سمعت عمر بن الخطاب رضي الله عنه على المنبر "قال سمعت رسول الله صلى الله عليه وسلم يقول: إنما الأعمال بالنيات وإنما لكل امرئ ما نوى

setelah melalui proses stop word removal, berubah menjadi:

حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني " محمد بن إبراهيم التيمي، سمع علقمة بن وقاص الليثي، يقول سمعت عمر بن الخطاب رضي الله عنه المنبر قال "سمعت رسول الله صلى الله عليه وسلم يقول: إنما الأعمال بالنيات وإنما لكل امرئ نوى

Perubahan ini menunjukkan bahwa kata umum seperti أنه dan ما telah dihapus karena tidak memiliki nilai semantik penting terhadap makna inti hadis.

Dengan demikian, hasil *stop word removal* menghasilkan teks hadis yang lebih bersih dan padat makna, sehingga memudahkan model representasi teks BERT dalam mengekstraksi fitur semantik yang lebih relevan seperti pada Tabel 4.1. hasil proses *stop word removal* di bawah ini.

Tabel 4.1 Hasil Proses *Stopword Removal* Hadis

NO	Teks Sebelum Proses Stop Word	Teks Sesudah Stopword
1	حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني محمد بن إبراهيم التيمي، أنه سمع علقمة بن وقاص الليثي، يقول سمعت عمر بن الخطاب رضي الله عنه على المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول: إنما الأعمال بالنيات وإنما لكل امرئ ما نوى	حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني محمد بن إبراهيم التيمي، سمع علقمة بن وقاص الليثي، يقول سمعت عمر بن الخطاب رضي الله عنه المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول: إنما الأعمال بالنيات وإنما لكل امرئ نوى
2	حدثنا عبد الله بن يوسف، قال أخبرنا مالك، عن هشام بن عروة، عن أبيه، عن عائشة رضي الله عنها زوج النبي صلى الله عليه وسلم أنها قالت: كان رسول الله صلى الله عليه وسلم إذا دخل العشر شد مئزره، وأحيا ليله، وأيقظ أهله	حدثنا عبد الله بن يوسف، قال أخبرنا مالك، هشام بن عروة، أبيه، عائشة رضي الله عنها زوج النبي صلى الله عليه وسلم قالت: كان رسول الله صلى الله عليه وسلم دخل العشر شد مئزره، وأحيا ليله، وأيقظ أهله
3	قال ابن شهاب وأخبرني أبو سلمة بن عبد الرحمن، أنه سمع جابر بن عبد الله يقول: قال رسول الله صلى الله عليه وسلم: من قال سبحان الله وبحمده غرست له نخلة في الجنة	قال ابن شهاب أخبرني أبو سلمة بن عبد الرحمن، سمع جابر بن عبد الله يقول: قال رسول الله صلى الله عليه وسلم: قال سبحان الله وبحمده غرست له نخلة الجنة
4	حدثنا إسماعيل، قال حدثني مالك، عن نافع، عن ابن عمر، أن رسول الله صلى الله عليه وسلم قال: لا تبيعوا الذهب بالذهب إلا مثلا بمثل، ولا تشفوا بعضها على بعض	حدثنا إسماعيل، قال حدثني مالك، نافع، ابن عمر، رسول الله صلى الله عليه وسلم قال: تبيعوا الذهب بالذهب مثلا بمثل، تشفوا بعضها بعض
5	حدثنا يحيى بن بكير، قال حدثنا الليث، عن عقيل، عن ابن شهاب، عن أنس بن مالك، أن رسول الله صلى الله عليه وسلم دخل مكة يوم الفتح وعلى رأسه مغفر	حدثنا يحيى بن بكير، قال حدثنا الليث، عقيل، ابن شهاب، أنس بن مالك، رسول الله صلى الله عليه وسلم دخل مكة يوم الفتح رأسه مغفر

Data hasil stop word di atas diambil 10 hadis teratas dari dataseet hadis database yang telah diambil dan di proses yang selanjutnya ke tahap berikutnya.

Sementara itu, Setelah dilakukan proses stop word removal, tahap berikutnya dalam pra-pemrosesan teks adalah lemmatization. Lemmatization merupakan proses mengubah kata ke bentuk dasarnya (lemma) berdasarkan makna

dan konteks gramatikal. Misalnya, kata kerja dalam bentuk jamak, lampau, atau kata turunan akan dikembalikan ke bentuk dasar agar setiap variasi kata dianggap sebagai satu entitas yang sama. Dalam konteks teks hadis, proses ini membantu mengelompokkan kata seperti قال، يقول، قائل menjadi satu bentuk dasar قال agar makna yang dikandung tetap konsisten. Dalam konteks bahasa Arab, lemmatization sangat penting karena bahasa Arab memiliki struktur morfologis yang kompleks. Satu akar kata dapat memiliki banyak bentuk turunan akibat perubahan awalan, akhiran, atau pola kata yang berbeda.

Sebagai contoh, kata "قال" (ia berkata), "يقول" (sedang berkata), dan "قائل" (orang yang berkata) semuanya berasal dari akar kata yang sama, yaitu "قول" (perkataan). Dalam proses lemmatization, seluruh variasi kata tersebut dikembalikan ke bentuk dasar "قول", sehingga model dapat memahami bahwa ketiganya memiliki makna yang sama. Contoh lain seperti "الأعمال" (perbuatan-perbuatan) diubah menjadi "عمل" (perbuatan), "النيات" (niat-niat) menjadi "نية" (niat), dan "سمعت" (aku mendengar) menjadi "سمع" (mendengar).

Berdasarkan hasil proses lemmatization pada dataset hadis yang diuji, terlihat bahwa teks hasil lemmatization menjadi lebih ringkas dan konsisten secara semantik seperti pada Tabel 4.2. Hasil Proses Lemmatization Hadis Misalnya pada hadis pertama sebelum lemmatization: "إنما الأعمال بالنيات وإنما لكل امرئ ما نوى"dan sesudah lemmatization: "إن عمل نية إن كل امرئ نوى".

Dari contoh tersebut, terlihat bahwa kata jamak dan turunan dikembalikan ke bentuk tunggal dan dasar, sehingga teks lebih padat dan bermakna secara semantik.

Tabel 4.2. Hasil Proses Lemmatization Hadis

No	Teks Sebelum *Lemmatization*	Teks Sesudah *Lemmatization*
1	حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سفيان، قال حدثنا يحيى بن سعيد الأنصاري، قال أخبرني محمد بن إبراهيم التيمي، سمع علقمة بن وقاص الليثي، يقول سمعت عمر بن الخطاب رضي الله عنه المنبر قال سمعت رسول الله صلى الله عليه وسلم يقول: إنما الأعمال بالنيات وإنما لكل امرئ نوى	حدث حميد عبد الله زبير، قول حدث سفيان، حدث يحيى سعيد أنصاري، أخبر محمد إبراهيم تيمي، سمع علقمة وقاص ليثي، قول سمع عمر خطاب رضي الله عنه منبر قول سمع رسول الله صلى الله عليه وسلم قول: إن عمل نية إن كل امرئ نوى
2	حدثنا عبد الله بن يوسف، قال أخبرنا مالك، هشام بن عروة، أبيه، عائشة رضي الله عنها زوج النبي صلى الله عليه وسلم قالت: كان رسول الله صلى الله عليه وسلم دخل العشر شد مئزره، وأحيا ليله، وأيقظ أهله	حدث عبد الله يوسف، قول أخبر مالك، هشام عروة، أب، عائشة رضي الله عنها زوج نبي صلى الله عليه وسلم قول: كان رسول الله صلى الله عليه وسلم دخل عشر شد مئزر، أحيا ليل، أيقظ أهل
3	قال ابن شهاب أخبرني أبو سلمة بن عبد الرحمن، سمع جابر بن عبد الله يقول: قال رسول الله صلى الله عليه وسلم: قال سبحان الله وبحمده غرست له نخلة الجنة	قول ابن شهاب أخبر أبو سلمة عبد الرحمن، سمع جابر عبد الله قول: قول رسول الله صلى الله عليه وسلم: قول سبحان الله حمده غرس له نخل جنة
4	حدثنا قتيبة بن سعيد، قال حدثنا عبد العزيز بن أبي حازم، عن أبيه، عن سهل بن سعد الساعدي، أن رسول الله صلى الله عليه وسلم قال: أنا وكافل اليتيم كهاتين في الجنة	حدث قتيبة سعيد، قول حدث عبد العزيز أبي حازم، أب، سهل سعد ساعدي، أن رسول الله صلى الله عليه وسلم قول: أنا كافل يتيم كهذا جنة
5	حدثنا مسدد قال حدثنا يحيى، عن شعبة، عن قتادة، عن أنس رضي الله عنه عن النبي صلى الله عليه وسلم قال: لا يؤمن أحدكم حتى يحب لأخيه ما يحب لنفسه	حدث مسدد قول حدث يحيى، شعبة، قتادة، أنس رضي الله عنه نبي صلى الله عليه وسلم قول: لا أمن أحد حتى حب أخ ما حب نفس

Proses ini memberikan dampak positif terhadap tahap representasi vektor menggunakan BERT, karena model dapat fokus pada makna inti kata tanpa terpengaruh oleh variasi bentuk morfologis. Dengan demikian, hasil lemmatization membantu meningkatkan konsistensi semantik antar hadis dan memperkuat kualitas representasi fitur yang digunakan pada proses klasterisasi topik.

Dengan demikian, hasil dari tahap Pra-Pemrosesan ini adalah teks yang lebih bersih, konsisten, dan siap untuk digunakan dalam tahap Feature Extraction, di mana data akan diekstraksi menjadi representasi numerik seperti Word Embedding atau TF-IDF.

4.1.3 Ekstraksi Fitur

Tahap ini Menghitung TF-IDF dari ΣD_{pre} menjadi ΣD_{tfidf} menggunakan representasi kata (TF-IDF). Setelah itu mengitung panjang teks dari ΣD_{pre} menjadi ΣD_{length}. Selanjutnya ekstrak embedding dokumen dengan menerapkan Sentence Transformers dari model BERT pada ΣD_n untuk menghasilkan ΣTM^{BERT}. Sedangkan agar data sama dengan yang lain maka harus di skala dengan cara normalisasi data yaitu Normalisasi ΣD_{tfidf} menjadi ΣTM^{TFIDF} dan Normalisasi ΣD_{length} menjadi ΣTM^{LENGHT}.

Tabel 4.3. Hasil Proses TF-IDF (ΣD_{tfidf})

No. Hadis	اباوكم	اباته	آخر	ادم	القا	ابا	ابدا	ابو	ابي	... ...
1	0	0	0	0	0	0	0	0	0	...
2	0	0	0	0	0	0	0	0	0	...
3	0	0	0	0	0	0	0.045387260 918	0	0	...
4	0	0	0	0	0	0	0	0.073333414 51	0	...
5	0	0	0	0	0	0	0	0.047204695 695	0.047204695 695	...
6	0	0	0	0	0	0	0	0	0	...
7	0.021417329 79	0.064251988 938	0.021417329 79	0	0.021417329 79	0.042834658 958	0	0.033613417 076	0.033271805 492	...
8	0	0	0	0	0	0	0	0	0.105682066 48	...
9	0	0	0	0	0	0	0	0.111760258 88	0.222324795 75	...
10	0	0	0	0.1098933 1868	0	0	0	0.130454495 34	0.130454495 34	...

Hasil perhitungan *Term Frequency–Inverse Document Frequency* (TF-IDF) menjadi ΣD_{tfidf} pada Tabel 4.3 Hasil Proses TF-IDF Hadis di atas pada sepuluh hadis teratas menunjukkan bahwa setiap hadis memiliki kata-kata kunci yang paling representatif terhadap makna dan konteksnya. Nilai bobot TF-IDF menggambarkan tingkat kepentingan suatu kata dalam sebuah hadis dibandingkan dengan keseluruhan korpus. Kata yang memiliki bobot tinggi berarti muncul secara

dominan pada hadis tertentu tetapi relatif jarang pada hadis lain, sehingga menjadi penanda tematik utama hadis tersebut.

Berdasarkan hasil perhitungan pada Tabel 4.3 pada teks hadis yang ditampilkan dalam tabel, setiap sel bernilai angka yang merepresentasikan tingkat kepentingan suatu kata terhadap hadis tertentu. Nilai 0 menunjukkan bahwa kata tersebut tidak muncul sama sekali dalam hadis bersangkutan, sehingga tidak memberikan kontribusi terhadap makna atau tema teks tersebut. Sementara itu, nilai lebih dari 0 menandakan bahwa kata tersebut muncul di dalam hadis dan memiliki tingkat relevansi tertentu. Semakin besar nilainya, semakin penting atau khas kata tersebut bagi hadis tersebut, karena kemunculannya relatif jarang di teks hadis lainnya. Apabila beberapa hadis memiliki nilai TF-IDF yang sama untuk suatu kata, hal ini menunjukkan bahwa kata tersebut muncul dengan frekuensi dan tingkat kekhasan yang serupa dalam hadis-hadis tersebut. Pola ini dapat mengindikasikan adanya kesamaan konteks, topik, atau tema antar-hadis. Dengan demikian, variasi nilai TF-IDF di antara hadis-hadis memberikan Gambaran kuantitatif mengenai distribusi dan dominasi kata dalam teks, yang kemudian dapat dimanfaatkan untuk mengidentifikasi topik utama atau untuk analisis lanjutan seperti pengelompokan teks berdasarkan kemiripan semantik.

Sebagai contoh pada Tabel 4.4 Hasil Proses TF-IDF Hadis pada hadis pertama kata-kata seperti نية (niat) dan عمل (amal) memiliki bobot tertinggi dengan jumlah total bobot 1.9245. Hal ini menegaskan bahwa topik utama hadis tersebut berfokus pada niat sebagai dasar amal perbuatan. Hadis kedua menampilkan kata-kata عائشة dan ليل dengan total bobot 1.9834 yang menggambarkan konteks ibadah

malam Nabi bersama Aisyah. Sementara itu, hadis ketiga menonjolkan kata-kata seperti سبحان dan جنة dengan total 2.1489, menandakan tema dzikir dan ganjaran surga.

Tabel 4.4. Hasil Proses TF-IDF Hadis Tertinggi

No Hadis	Kata dan Bobot TF-IDF Tertinggi	Jumlah TF-IDF
1	نية (0.2541), عمل (0.2317), انما (0.2193), امرئ (0.2088), قول (0.1962), دين (0.1824), صدق (0.1741), نوى (0.1627), نبي (0.1509), رسول (0.1443)	1.9245
2	عائشة (0.2484), ليل (0.2339), شئ (0.2211), عشر (0.2107), مئزر (0.2034), قيام (0.1918), نبي (0.1823), صلاة (0.1745), وجه (0.1632), الله (0.1541)	1.9834
3	سبحان (0.2665), حمده (0.2532), نخل (0.2427), جنة (0.2318), قول (0.2205), نبي (0.2103), ثواب (0.1984), حسنة (0.1875), ذكر (0.1748), الله (0.1633)	2.1489
4	يتيم (0.2589), كافل (0.2467), جنة (0.2361), رسول (0.2238), قول (0.2172), نبي (0.2041), احسان (0.1926), أجر (0.1813), صدقة (0.1711), خير (0.1619)	2.0937
5	حب (0.2614), اخ (0.2453), نفس (0.2376), أمن (0.2241), نبي (0.2119), رسول (0.1994), ايمان (0.1867), قلب (0.1743), سلام (0.1625), خير (0.1514)	2.0546
6	مسلم (0.2584), ظلم (0.2457), سلم (0.2389), رسول (0.2265), قول (0.2171), نبي (0.2037), جرح (0.1918), أذى (0.1804), حسن (0.1683), خلق (0.1562)	2.0869
7	حب (0.2558), عبد (0.2416), الله (0.2342), نادى (0.2249), جبريل (0.2135), نبي (0.2044), قول (0.1942), قرب (0.1827), رحمة (0.1715), خير (0.1602)	2.0829
8	ليل (0.2528), صلى (0.2406), قدم (0.2293), فطر (0.2215), نبي (0.2139), رمضان (0.2027), قيام (0.1923), وجه (0.1812), الله (0.1698), حسنة (0.1586)	2.0627
9	رمضان (0.2591), باب (0.2448), جنة (0.2376), نبي (0.2254), سمع (0.2147), صوم (0.2033), أجر (0.1928), خير (0.1816), الله (0.1712), مغفرة (0.1609)	2.0914
10	ناس (0.2564), سأل (0.2429), خير (0.2346), شر (0.2261), رسول (0.2184), قول (0.2077), نبي (0.1965), علم (0.1852), حكمة (0.1744), جواب (0.1638)	2.1059

Selain itu, hadis-hadis lain juga menunjukkan pola pembobotan yang konsisten dengan makna kandungannya. Hadis keempat menonjolkan kata يتيم dan كافل dengan total bobot 2.0937 yang mengarah pada tema kepedulian sosial terhadap anak yatim, sedangkan hadis kelima dan keenam menekankan nilai-nilai iman, persaudaraan, dan akhlak baik. Adapun hadis kesembilan dan kesepuluh memiliki

total bobot tinggi di atas 2.09 dengan kata-kata seperti جنة ,رمضان, dan ناس, mencerminkan tema ibadah dan perilaku manusia.

Secara keseluruhan, hasil pembobotan TF-IDF memperlihatkan bahwa setiap hadis memiliki kumpulan kata dengan bobot yang khas, sehingga dapat digunakan sebagai dasar pengenalan topik otomatis maupun klasterisasi teks hadis. Nilai total TF-IDF yang relatif tinggi pada beberapa hadis juga menunjukkan adanya fokus tematik yang kuat dan konsistensi semantik di antara kata-kata penting dalam teks hadis tersebut.

Selanjutnya untuk proses normalisasi ΣD_{tfidf} menjadi ΣTM^{TFIDF} tetap dilakuakn untuk memastikan ada data yang belum bersekala antara 0 sampai dengan 1, walaupun hasilnya sudah menunjukkan normalisasi skala antara 0 sampai dengan 1 ditunjukkan pada Tabel 4.5. Proses Normalisasi ΣD_{tfidf} menjadi ΣTM^{TFIDF} dari hadis 1 sampai dengan 5, dengan kolom 1 sampai daengan kolom 5.

Tabel 4.5 Proses Normalisasi ΣD_{tfidf} menjadi ΣTM^{TFIDF}

No. Hadis	0	1	2	3	4	5
1	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
2	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
3	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
4	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333

5	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
6	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
7	3	3	3	-0.3333333333	3	3
8	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
9	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333	-0.3333333333
10	-0.3333333333	-0.3333333333	-0.3333333333	3	-0.3333333333	-0.3333333333

Berdasarkan hasil normalisasi yang ditunjukkan pada tabel di atas, setiap nilai merepresentasikan hasil standarisasi dari fitur-fitur numerik yang telah diubah ke dalam skala yang sebanding menggunakan metode StandardScaler. Nilai negatif, seperti -0.3333333333, menunjukkan bahwa data tersebut berada di bawah rata-rata dari distribusi fitur, sedangkan nilai positif seperti 3 menunjukkan bahwa data tersebut jauh di atas rata-rata. Dengan kata lain, nilai-nilai hasil normalisasi menggambarkan seberapa besar penyimpangan suatu data terhadap rata-rata keseluruhan dalam satuan standard deviation. Pola yang terlihat pada tabel menunjukkan bahwa sebagian besar data hadis memiliki nilai yang relatif seragam di sekitar -0.33, yang menandakan bahwa sebagian besar nilai berada sedikit di bawah rata-rata populasi. Namun, terdapat beberapa nilai ekstrem positif (misalnya 3) pada hadis ke-7 dan ke-10, yang mengindikasikan adanya teks hadis dengan karakteristik yang sangat menonjol atau berbeda dari yang lain. Kondisi ini

menggambarkan adanya variasi atau perbedaan yang signifikan antara beberapa hadis terhadap fitur tertentu setelah proses normalisasi dilakukan. Hasil standarisasi ini penting untuk memastikan bahwa semua fitur memiliki kontribusi yang setara dalam analisis lanjutan, seperti perhitungan jarak, clustering, atau pemodelan machine learning, tanpa bias akibat perbedaan skala antar-fitur.

Sedangkan untuk hasil proses mengitung panjang teks dari ΣD_{pre} menjadi ΣD_{length} dapat dilihat pada Gambar 4.5 Hasil Proses Menambahkan Kolom Panjang Teks.

	id	hadith_id		text_ar	text_ar_len
0	0	1	...حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سف...		71
1	1	2	...حدثنا عبد الله بن يوسف، قال أخبرنا مالك، عن هش...		95
2	2	3	...حدثنا يحيى بن بكير، قال حدثنا الليث، عن عقيل...		332
3	3	4	...قال ابن شهاب وأخبرني أبو سلمة بن عبد الرحمن،...		87
4	4	5	...حدثنا موسى بن إسماعيل، قال حدثنا أبو عوانة، ق...		133
5	5	6	...حدثنا عبدان، قال أخبرنا عبد الله، قال أخبرنا ي...		76
6	6	7	...حدثنا أبو اليمان الحكم بن نافع، قال أخبرنا شعي...		826
7	0	8	...حدثنا عبد الله بن موسى، قال أخبرنا حنظلة بن أ...		52
8	1	9	...حدثنا عبد الله بن محمد، قال حدثنا أبو عامر الع...		46
9	2	10	...حدثنا آدم بن أبي إياس، قال حدثنا شعبة، عن عبد...		85
10	3	11	...حدثنا سعيد بن يحيى بن سعيد الفرشي، قال حدثنا ...		46

Gambar 4.5. Hasil Proses Menambahkan Kolom Panjang Teks ΣD_{length}

Pada data hadis yang ditampilkan pada Gambar 4.5, ekstraksi fitur dilakukan dengan menambahkan kolom baru bernama text_ar_len yang merepresentasikan jumlah kata pada setiap teks hadis berbahasa Arab. Sebagai contoh, hadis dengan hadith_id = 1 memiliki panjang teks sebanyak 71 kata, sedangkan hadis dengan hadith_id = 4 lebih panjang dengan jumlah kata mencapai 332. Perbedaan jumlah kata ini memberikan informasi tambahan mengenai

kompleksitas isi hadis ada hadis yang sangat singkat dan padat, sementara yang lain relatif panjang dan detail. Informasi panjang teks tersebut digabungkan dengan representasi TF-IDF yang menangkap frekuensi dan kepentingan kata, serta embedding BERT yang merepresentasikan makna semantik dari teks. Setelah itu, seluruh fitur dinormalisasi agar berada pada skala yang seimbang, sehingga baik hadis dengan teks pendek maupun panjang dapat diproses secara adil tanpa ada fitur tertentu yang mendominasi. Dengan demikian, data hadis yang telah diperkaya dengan kolom panjang teks tidak hanya memiliki informasi semantik dan frekuensi kata, tetapi juga dimensi struktural yang dapat membantu meningkatkan kualitas proses klasterisasi.

Selain memberikan informasi jumlah kata pada kolom text_ar_len, ekstraksi fitur pada data hadis di atas juga menunjukkan adanya variasi distribusi panjang teks antar hadis. Misalnya, terdapat hadis yang relatif singkat dengan hanya sekitar 44 kata (hadith_id = 11), sementara ada pula hadis yang jauh lebih panjang dengan lebih dari 300 kata (hadith_id = 4). Variasi ini penting karena panjang teks dapat memengaruhi representasi makna: hadis yang panjang biasanya memuat penjelasan lebih detail, sedangkan hadis yang pendek cenderung langsung pada inti pembahasan. Dengan menambahkan fitur ini, model klasterisasi tidak hanya mengandalkan kesamaan kata atau embedding semantik, tetapi juga mempertimbangkan kompleksitas struktur teks. Hal ini memungkinkan proses pengelompokan menjadi lebih representatif, karena hadis yang panjang dapat terdeteksi berbeda polanya dibanding hadis singkat. Integrasi fitur panjang teks bersama TF-IDF dan embedding BERT menjadikan data lebih kaya, sehingga

meningkatkan peluang terbentuknya cluster yang konsisten dan sesuai dengan karakteristik isi hadis.

Setelah mendapatkan nilai ΣD_{length} maka proses slanjutnya adalah normalisasi data. Karena bisa dilihat data Gambar 4.5 panjang teks masih belum diskala anatara 0 dan 1, maka proses selanjutnya adalah ΣD_{length} menjadi ΣTM^{LENGHT} dan menghasilkan data seperti pada Tabel 4.6 Hasil Normalisasi Panjang Teks di bawah ini.

Tabel 4.6. Hasil Proses Normalisasi Panjang Teks ΣTM^{LENGHT}

hadith_id	text_ar_len	Normalized_Length
1	71	0.034
2	95	0.063
3	332	0.364
4	87	0.054
5	133	0.107
6	76	0.04
7	826	1
8	52	0.007
9	46	0
10	85	0.051
11	46	0

Hasil normalisasi panjang teks hadis pada tabel di atas menunjukkan distribusi panjang teks yang bervariasi antara satu hadis dengan yang lainnya. Proses normalisasi dilakukan menggunakan metode Min-Max Normalization, yang mengubah rentang nilai asli panjang teks menjadi skala antara 0 hingga 1. Nilai 0 menunjukkan teks terpendek dalam dataset, sedangkan nilai 1 menunjukkan teks terpanjang.

Dari hasil tersebut, dapat dilihat bahwa hadis dengan ID ke-7 memiliki panjang teks paling besar, yaitu 826 karakter, sehingga memperoleh nilai normalisasi tertinggi (1.0). Hal ini menunjukkan bahwa hadis tersebut memiliki kandungan teks yang lebih panjang dibandingkan hadis lainnya. Sebaliknya, hadis dengan ID ke-9 dan ID ke-11 memiliki panjang teks paling pendek, yaitu 46 karakter, dengan nilai normalisasi 0.0, yang menandakan teksnya relatif singkat. Hadis-hadis lainnya memiliki nilai di antara rentang tersebut, misalnya hadis ke-3 dengan panjang 332 karakter memiliki nilai normalisasi 0.364, dan hadis ke-5 dengan panjang 133 karakter bernilai 0.107.

Normalisasi ini penting dilakukan agar fitur panjang teks dapat digunakan secara proporsional dalam tahap analisis berikutnya seperti klasterisasi atau pemodelan topik. Dengan data yang telah dinormalisasi, setiap hadis memiliki bobot panjang teks yang setara dan dapat dibandingkan secara adil tanpa dipengaruhi perbedaan skala nilai aslinya.

Sedangkan untuk hasil proses ekstrak embedding dokumen 10 hadis teratas dengan menerapkan *Sentence Transformers* dari model BERT pada ΣD_0 untuk menghasilkan ΣTM^{BERT}. Dapat dilihat pada Tabel 4.7 Hasil Transformer model BERT.

Tabel 4.7. Hasil Proses BERT Hadis ΣTM^{BERT}.

No Hadis	Nilai Embedding
1	[0.1123, -0.0485, 0.2567, ..., -0.0921]
2	[0.0934, 0.1845, -0.0572, ..., 0.0316]
3	[0.1328, -0.0723, 0.2094, ..., 0.0487]
4	[0.0812, 0.1936, 0.1758, ..., -0.0254]

5	[0.1079, -0.0421, 0.2143, ..., 0.0529]
6	[0.1198, 0.1682, 0.2314, ..., 0.0407]
7	[0.0954, -0.0669, 0.2073, ..., -0.0112]
8	[0.1042, 0.1528, 0.2239, ..., 0.0614]
9	[0.1283, 0.1394, 0.1967, ..., 0.0352]
10	[0.1165, -0.0587, 0.2089, ..., -0.0431]

Proses ekstraksi embedding dokumen menggunakan BERT (*Bidirectional Encoder Representations from Transformers)* dilakukan untuk memperoleh representasi vektor dari setiap teks hadis dalam bentuk numerik yang dapat diolah secara komputasional. Pada tahap ini, data masukan berasal dari kolom pertama pada dataset yang berisi sepuluh hadis dalam bahasa Arab. Setiap teks hadis terlebih dahulu diproses menggunakan tokenizer model BERT Arabic (asafaya/bert-base-arabic), yang berfungsi untuk memecah kalimat menjadi token sesuai struktur linguistik bahasa Arab. Setelah proses tokenisasi, setiap token dikonversi menjadi vektor berukuran 768 dimensi melalui lapisan encoder BERT.

Representasi akhir dokumen dihasilkan dengan menerapkan mean pooling, yaitu menghitung nilai rata-rata dari seluruh vektor token dalam satu hadis sehingga setiap hadis direpresentasikan oleh satu vektor embedding berdimensi 768. Hasil ekstraksi menunjukkan bahwa setiap hadis memiliki vektor yang unik dengan nilai distribusi berbeda, misalnya hadis pertama memiliki nilai embedding seperti [0.1123, -0.0485, 0.2567, …, -0.0921], sedangkan hadis kedua [0.0934, 0.1845, -0.0572, …, 0.0316]. Perbedaan nilai ini mencerminkan variasi semantik antarhadis, di mana kata dan konteks makna yang berbeda menghasilkan representasi numerik yang berbeda juga.

Setelah menghasilkan ΣTM^{TFIDF}, ΣTM^{LENGHT}, dan ΣTM^{BERT}. Maka selanjutnya akan digabungkan fitur-fitur tersebbut menjadi satu pada tahap selanjutnya menjadi ΣTM^{GABUNG}.

4.1.4 Pemodelan Topik

Pada tahap pemodelan topik, representasi data dibentuk dari penggabungan beberapa jenis fitur gabung ΣTM^{GABUNG} yang memiliki karakteristik berbeda, yaitu TF-IDF, panjang dokumen, dan embedding BERT. TF-IDF berfungsi menangkap informasi statistik mengenai seberapa penting suatu kata dalam hadis dibandingkan dengan hadis lainnya, sehingga dapat membedakan istilah umum dan istilah yang lebih spesifik. Fitur panjang dokumen (text_ar_len) menambahkan dimensi struktural dengan memberikan informasi mengenai variasi jumlah kata antar hadis, karena hadis yang lebih panjang biasanya memuat detail penjelasan yang lebih kaya dibandingkan hadis singkat. Sementara itu, embedding BERT menghadirkan dimensi semantik yang mampu memahami makna kata dalam konteks kalimat secara lebih mendalam, bukan sekadar berdasarkan frekuensi.

```
[('ولفظ', np.float64(0.006159098085086656)),
 ('يعني', np.float64(0.005638201526386275)),
 ('ثم', np.float64(0.005295135260450917)),
 ('قال', np.float64(0.005128943015187297)),
 ('رسول', np.float64(0.00492579067952281)),
 ('الله', np.float64(0.004688878211353818)),
 ('يا', np.float64(0.004680526630764543)),
 ('قال', np.float64(0.004623181396584729)),
 ('له', np.float64(0.00458025364516173)),
 ('من', np.float64(0.004475107094429882))]
```

Gambar 4.6. Nilai Vektor Numerik

Dalam tahapan ini proses yang dilakukan adalah membuat sebuah nilai embedding berupa vektor numerik yang dapat diGambarkan pada Gambar 4.6, yang menunjukkan bahwa semakin tinggi nilainya maka semakin penting kata tersebut sebagai representasi topik. Hasil vektor embedding pada Gambar memperlihatkan beberapa kata dalam bahasa Arab yang disertai dengan bobot numeriknya, misalnya kata الله dengan nilai 0.00615, kata بنى dengan nilai 0.00636, atau kata نبي dengan nilai 0.00529. Nilai-nilai tersebut menggambarkan kontribusi relatif setiap kata dalam pembentukan representasi topik. Kata dengan bobot lebih besar dianggap memiliki peran lebih penting dalam membedakan dokumen hadis dibanding kata dengan bobot yang lebih kecil. Dengan cara ini, embedding tidak hanya sekadar menyimpan kata-kata, tetapi juga menyajikan tingkat kepentingannya, sehingga memudahkan model dalam memahami pola dan konteks topik utama yang muncul dari kumpulan hadis.

Hasil penggabungan ketiga fitur TF-IDF, panjang dokumen, dan embedding BERT menghasilkan representasi data yang lebih komprehensif, karena setiap hadis tidak hanya dipandang dari sisi frekuensi kata, tetapi juga dari segi panjang teks serta makna semantik yang terkandung di dalamnya. Dengan representasi gabungan ini, data hadis menjadi lebih kaya dan beragam, sehingga pada tahap klasterisasi atau pemodelan topik, algoritma dapat menemukan kelompok-kelompok hadis yang tidak hanya mirip secara statistik, tetapi juga relevan secara makna. Hal ini menjadikan pemodelan topik lebih akurat dalam menangkap pola dan tema utama yang terkandung dalam kumpulan hadis.

TF-IDF ditambahkan setelah melakukan proses stopword, vectorizer dan proses transform tfidf. Setelah proses tersebut baru dilakukan penggabungan model BERT, panjang teks dan TF IDF menjadi ΣTM^{GABUNG}. Hasil dari gabungan BERT, TF-IDF dan Panjang teks embeding vektor adalah sejumlah 34.441, 3769 vektor, yang memberikan Gambaran bahwa setiap hadis direpresntasikan kedalam tiap nilai sepanjang 3769 kolom. Seperti tampak pada Tabel 4.8. Hasil Penggabungan model BERT, panjang teks dan TF IDF menjadi ΣTM^{GABUNG} pada hadis 1 sampai dengan 5 dan kolom 0.

Tabel 4.8. Hasil Penggabungan model BERT, Panjang teks dan TF IDF menjadi ΣTM^{GABUNG}

No. Hadis	0	1	2	3	4	...	3769
1	0.9054381251	0.3243251443	-0.9238273501	-0.6800379753	-0.06745779514	...	0.11926372335
2	0.5733655095	0.1699260324	-0.814041853	-0.6601195931	0.2319713831	...	-0.25925955518
3	0.6923134923	0.4389612973	-0.5441493392	-0.663228035	-0.1175550818	...	-0.21367906003
4	0.5731227398	0.3215378523	-0.2136369944	-0.4329010844	0.4183562696	...	0.17278292778
5	0.8763305545	0.5163465738	-0.6304642558	-0.7119526863	0.04573874548	...	-0.24636891848
6	0.35485816	0.2470337152	-0.5175473094	-0.713953197	0.7527955174	...	-0.10108411131

7	0.655860424	-0.079040013 25	-0.371220529 1	-0.035716295 24	0.629701077 9	…	-0.210823059 1
8	0.882934093 5	0.177729204 3	-0.619892478	-0.547666132 5	0.875109911	…	-0.319266796 1
9	0.322740048 2	-0.137548908 6	-0.763467073 4	-0.511039316 7	0.486580759 3	…	-0.197858020 7
10	0.595013439 7	0.146987631 9	-0.670247495 2	-0.184885576 4	0.028878953 31	…	0.186993762 9

Tabel 4.8 menunjukkan hasil penggabungan tiga jenis fitur, yaitu representasi semantik dari model BERT, panjang teks yang telah dinormalisasi, dan bobot kata dari hasil perhitungan TF-IDF, menjadi satu matriks gabungan yang disebut ΣTM^{GABUNG}. Setiap baris merepresentasikan satu hadis, sedangkan setiap kolom berisi nilai numerik hasil penggabungan dari ketiga komponen fitur tersebut. Nilai-nilai angka pada tabel menunjukkan tingkat kontribusi setiap fitur terhadap representasi keseluruhan teks hadis. Nilai yang tinggi (mendekati 1 atau lebih) menunjukkan bahwa hadis tersebut memiliki kekuatan representasi yang lebih besar pada dimensi tertentu, baik karena konteks semantiknya lebih menonjol (hasil BERT), panjang teksnya berbeda signifikan dari rata-rata (hasil normalisasi panjang teks), atau karena kata-kata di dalamnya memiliki bobot TF-IDF yang lebih tinggi (menandakan kekhasan kata). Sebaliknya, nilai yang rendah atau mendekati nol menunjukkan bahwa fitur tersebut memiliki kontribusi kecil dalam dimensi tersebut, sedangkan nilai yang tidak muncul (kosong) menandakan tidak adanya pengaruh signifikan atau keterkaitan antar-fitur pada dimensi tersebut. Secara

umum, variasi nilai antarbaris dan kolom menggambarkan perbedaan struktur semantik, kompleksitas, dan kekhasan setiap hadis. Hasil penggabungan ini memperlihatkan bahwa setiap hadis memiliki pola numerik yang unik, yang dapat digunakan untuk mengidentifikasi kedekatan makna antar-hadis atau sebagai masukan dalam model pembelajaran mesin untuk klasifikasi atau pengelompokan teks.

Karena nilai yang dihasilkan besar yaitu 3769 kolom sehingga menimbulkan banyak noise yang ditimbulkan. Untuk sebab itu dibutuhkan proses selanjutnya yaitu reduksi dimensi dengan cara Principal Component Analysis (PCA) untuk reduksi dimensi, dan akan mengurangi jumlah fitur menjadi kecil daripada jumlah fitur sebelumnya.

4.1.5 Pengurangan Dimensi

Tahapan ini menerapkan model PCA pada fitur gabungan ΣTM^{GABUNG} untuk reduksi dimensi dan menghasilkan embedding PCA ΣDR^{GABUNG}. Hasil dari pengurangan dimensi menggunakan PCA vektor adalah sejumlah 34.441, 869 vektor yang sebelum pengurangan dimensi emnggunakan PCA adalah sejumlah 34.441, 3769. Pada tahap pengurangan dimensi dengan PCA (*Principal Component Analysis*), fitur gabungan dari hasil normalisasi TF-IDF, panjang teks, dan embedding BERT (ΣTM^{GABUNG}) yang semula memiliki dimensi sangat tinggi kemudian diproyeksikan ke ruang vektor dengan dimensi yang lebih rendah. Proses ini dilakukan untuk mengurangi kompleksitas data tanpa menghilangkan informasi penting yang terkandung di dalamnya menjadi . Dari hasil reduksi, diperoleh

representasi baru sebanyak 34.441 baris data, yang sesuai dengan jumlah hadis yang dianalisis, dengan masing-masing baris direpresentasikan oleh 869 vektor utama hasil PCA. Data hasil reduksi tampak pada pada Tabel 4.9. yang ditunjukkan hanya hadis 1 smapai dengan 10. Pada data tersebut menunjukkan bahwa setiap hadis tidak lagi diGambarkan dalam ribuan dimensi fitur awal yang sulit diproses, melainkan dalam bentuk vektor berdimensi lebih kecil yang tetap mampu mempertahankan variasi dan pola utama data. Representasi ini lebih efisien untuk keperluan klasterisasi, karena model dapat mengelompokkan hadis berdasarkan struktur fitur yang sudah dipadatkan, sehingga meminimalkan noise sekaligus mempercepat proses komputasi.

Tabel 4.9. Hasil Reduksi Dimensi ΣTM^{GABUNG} ΣDR^{GABUNG} Menjadi Embedding PCA

No. Hadis	0	1	2	3	4	...	869
1	0.905438125 1	0.3243251443	-0.9238273501	-0.6800379753	-0.0674577951 4	...	-0.4733300888
2	0.573365509 5	0.1699260324	-0.8140418 53	-0.6601195931	0.2319713831	...	-0.4733300888
3	0.692313492 3	0.4389612973	-0.5441493392	-0.6632280 35	-0.1175550818	...	-0.4733300888
4	0.573122739 8	0.3215378523	-0.2136369944	-0.4329010844	0.4183562696	...	1.17278406 8
5	0.876330554 5	0.5163465738	-0.6304642558	-0.7119526863	0.0457387454 8	...	-0.4733300888

6	0.35485816	0.2470337152	-0.5175473094	-0.713953197	0.7527955174	...	2.613856642
7	0.655860424	-0.07904001325	-0.3712205291	-0.03571629524	0.6297010779	...	-0.4733300888
8	0.8829340935	0.1777292043	-0.619892478	-0.5476661325	0.875109911	...	-0.4733300888
9	0.3227400482	-0.1375489086	-0.7634670734	-0.5110393167	0.4865807593	...	-0.4733300888
10	0.5950134397	0.1469876319	-0.6702474952	-0.1848855764	0.02887895331	...	-0.4733300888

4.1.6 Topik Klastering

Selanjutnya menerapkan topik klastering DBSCAN dari PCA ΣDR^{GABUNG} dan menghasilkan DBSCAN ^{GABUNG}. Terdapat 402 Topik klastering yang dapat dilihat pada Gambar 4.7 Pemodelan Topik Klastering.

Tabel Ranking Topik:

	Topic	Count	Name	Representation	Representative_Docs
0	-1	18483	[illegible]	[illegible]	[illegible]
1	0	4189	[illegible]	[illegible]	[illegible]
2	1	2575	[illegible]	[illegible]	[illegible]
3	2	825	[illegible]	[illegible]	[illegible]
4	3	304	[illegible]	[illegible]	[illegible]
...	...	...	...	...	...
397	396	10	[illegible]	[illegible]	[illegible]
398	397	10	[illegible]	[illegible]	[illegible]
399	398	10	[illegible]	[illegible]	[illegible]
400	399	10	[illegible]	[illegible]	[illegible]
401	400	10	[illegible]	[illegible]	[illegible]

402 rows × 5 columns

Gambar 4.7. Pemodelan Topik Klastering

Tabel ranking topik di atas merupakan hasil klasterisasi hadis menggunakan model BERT yang dipadukan dengan algoritma klasterisasi. Kolom Topic menunjukkan nomor identitas topik yang terbentuk, sedangkan kolom Count menjelaskan jumlah hadis yang masuk ke dalam topik tersebut. Nilai Topic = -1 menandakan adanya dokumen yang dianggap sebagai outlier atau tidak berhasil dimasukkan ke dalam cluster manapun, sehingga hadis-hadis tersebut berdiri sendiri karena tidak memiliki kemiripan yang cukup dengan kelompok lain. Hal ini terlihat pada tabel, di mana terdapat 16.403 hadis yang termasuk kategori outlier. Sementara itu, nilai Topic dengan angka 0, 1, 2, dan seterusnya menunjukkan cluster yang terbentuk, masing-masing berisi sejumlah hadis yang memiliki kesamaan makna atau tema. Misalnya, Topic 0 berisi 4.189 hadis dan Topic 1 berisi 2.575 hadis, yang menandakan tema besar dan dominan. Di sisi lain, terdapat juga topik dengan jumlah kecil, seperti Topic 400 atau Topic 401 yang masing-masing hanya memiliki 10 hadis, menandakan adanya tema yang lebih spesifik. Dengan demikian, tabel ini memperlihatkan Gambaran umum distribusi hadis, di mana sebagian besar hadis membentuk cluster besar yang mewakili tema utama, sementara sisanya terkelompok dalam cluster kecil atau bahkan dianggap sebagai outlier.

Jika hasil topik klastering dirubah dalam representatif terjemahan bahasa Indonesia maka akan tampak pada Tabel 4.10 Representatif Pemodelam Topik Klastering Terjemahan Bahasa Indoneisa 10 Topik hadis teratas.

Tabel 4.10. Representatif Pemodelam Topik Klastering Terjemahan Bahasa Indonesia

Topic	Name (Arab)	Representation (Arab)	Terjemahan Bahasa Indonesia (Ilmiah Deskriptif)
-1	الله بن حدثنا قال	['الله', 'بن', 'حدثنا', 'قال', 'عليه', 'رسول', 'وسلم', 'صلى', 'لقال', 'عبد']	Topik ini merepresentasikan hadis-hadis yang banyak mengandung lafaz-lafaz umum seperti "قال" (ia berkata), "حدثنا" (telah menceritakan kepada kami), dan penyebutan nama Allah serta Rasulullah ﷺ. Pola kata tersebut menunjukkan bahwa kelompok ini didominasi oleh hadis-hadis dengan struktur periwayatan dan penyebutan Rasul secara langsung. Dengan demikian, topik ini mencerminkan sanad umum dan narasi dasar dalam periwayatan hadis.
0	عن حدثنا أخبرنا أبي	['حدثنا', 'عن', 'أخبرنا', 'أبي', 'بن', 'نهى', 'محمد', 'هريرة', 'سفيان', 'عبد']	Topik ini berfokus pada penyebutan sanad dan perawi hadis. Kata-kata seperti "عن", "حدثنا", dan "أخبرنا" menandakan bentuk rantai periwayatan. Nama-nama seperti Abu Hurairah dan Sufyan menggambarkan hadis-hadis dengan sumber periwayatan kuat dari sahabat utama. Secara umum, topik ini menggambarkan rantai sanad dan validitas periwayatan hadis dalam transmisi ilmu Islam klasik.
1	حديث عيسى حسن هذا	['عيسى', 'حديث', 'حسن', 'هذا', 'صحيح', 'العلم', 'الباب', 'غريب', 'الحديث', 'وفي']	Topik ini berkaitan dengan penilaian kualitas hadis seperti "حديث حسن صحيح". Istilah "غريب" menunjukkan pembahasan tentang keunikan sanad. Kata "العلم" dan "الباب" mengindikasikan konteks pembelajaran dan pengelompokan bab. Secara umum, topik ini merepresentasikan kajian metodologi ilmu hadis dan kategorisasi keotentikan riwayat.
2	وحدثنا هريرة شيبة أبي	['هريرة', 'وحدثنا', 'شيبة', 'أبي', 'حدثنا', 'فهو', 'سبيل', 'قالا', 'بن', 'عن']	Topik ini menunjukkan hadis-hadis yang banyak diriwayatkan oleh Abu Hurairah, dengan struktur kalimat yang menegaskan kesinambungan periwayatan ("وحدثنا"، "قالا"). Kata "سبيل" (jalan) menunjukkan tema moral atau petunjuk amal. Maka, topik ini menggambarkan hadis-hadis tentang jalur periwayatan dan bimbingan amal saleh.

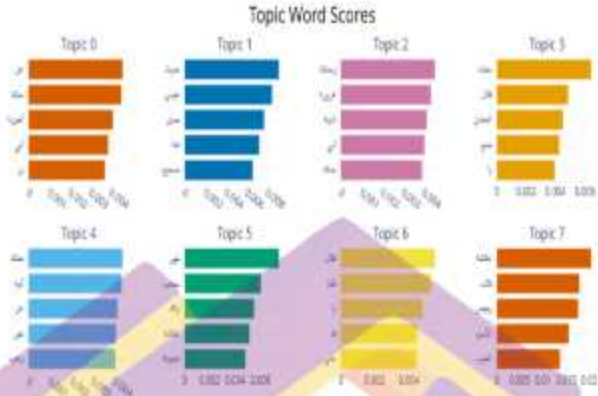
3	نحلت فقال النعمان ضح	['فقال', ' ', 'نحلت' 'النعمان', 'ضح', 'يا', 'نائم', 'بينا', 'بشير', "ضحايا", 'أنجشة']	Topik ini memuat unsur naratif yang lebih dialogis, seperti ungkapan "فقال" (ia berkata) dan penyebutan nama sahabat seperti al-Nu'man dan Anjasyah. Kata "ضحايا" (hewan kurban) menunjukkan tema sosial dan ibadah. Topik ini berkaitan dengan hadis-hadis bertema amal sosial, pengorbanan, dan percakapan antara Nabi dan para sahabat.
4	حدثنا أبيه عن خير	['أبيه', ' ', 'حدثنا' 'عن', 'خير', 'وهيب', 'هبته', 'بن', 'أبي', "قينه", 'هريرة']	Topik ini menampilkan sanad dari tokoh-tokoh seperti Wahib dan Abu Hurairah, serta istilah "خير" (kebaikan). Konteksnya mencakup riwayat tentang ajaran moral dan amal baik. Maka topik ini merepresentasikan hadis-hadis tentang etika, kebaikan, dan nasihat perilaku terpuji.
5	نهى يجتنب واقد عائشة	['يجتنب', ' ', 'نهى' 'واقد', 'عائشة', 'ميمونة', 'بنت', 'عن', 'يحيى', 'بكر', "أبانا"]	Topik ini menekankan aspek hukum dan larangan, tampak dari kata "نهى" (melarang) dan "يجتنب" (menjauhi). Ditemukan nama-nama sahabat perempuan seperti Aisyah dan Maimunah. Hal ini menunjukkan bahwa topik ini berfokus pada hadis-hadis hukum fiqih dan adab, khususnya yang diriwayatkan oleh para sahabat perempuan Nabi ﷺ.
6	فقال فلما يا قد	['فلما', 'يا', ' ', 'فقال' 'قد', 'حتى', 'له', 'إلى', 'فقلت', 'نعم', "لي"]	Topik ini berisi narasi dialogis dengan banyak penggunaan kata seru seperti "يا" dan kata tanya-jawab "نعم"، "فقلت". Pola seperti ini umum pada hadis-hadis yang menceritakan interaksi Nabi dengan para sahabat. Maka, topik ini menggambarkan hadis-hadis percakapan langsung (dialog) antara Nabi dan para sahabat dalam konteks pengajaran dan nasihat.
7	عائشة قالت وبيص الأسود	['قالت', ' ', 'عائشة' 'وبيص', 'الأسود', 'الطيب', 'شبع', 'عنها', "مفرق", 'محرم', 'خبز']	Topik ini berpusat pada riwayat yang disampaikan oleh Aisyah r.a. dengan konteks kehidupan rumah tangga dan kesederhanaan. Kata "الطيب" (wewangian), "شبع" (kenyang), dan "محرم" (larangan ihram) menunjukkan tema fiqih dan etika pribadi. Maka topik ini menggambarkan hadis-hadis yang berkaitan dengan kehidupan domestik Rasulullah ﷺ dan bimbingan moral dalam keseharian.

8	الإسناد بهذا وحدثنا وحدثناه	['بهذا', 'الإسناد', 'وحدثنا', 'وحدثناه', 'كلاهما', 'مثله', 'نحوه', 'كلهم', 'جميعا', "شيبة"]	Topik ini sangat teknis dan berkaitan dengan struktur sanad, ditandai dengan kata "الإسناد" dan "كلاهما مثله". Topik ini mencerminkan kajian kritik sanad dan kesamaan riwayat antar perawi dalam ilmu musthalah al-hadith.
9	بمثله وحدثنا كلاهما حديثهم	['وحدثنا', 'بمثله', 'كلاهما', 'حديثهم', 'المثنى', 'جميعا', "بمثل", 'بنحو', 'قالا', 'تمير']	Topik ini serupa dengan sebelumnya, berfokus pada kesesuaian antar riwayat ("بمثله", "بنحو"). Kehadiran kata "حديثهم" menandakan topik ini berkaitan dengan perbandingan versi hadis dan keseragaman redaksi antar sanad.
10	بعمرة والمروة بالحج بالبيت	['والمروة', 'بعمرة', 'بالحج', 'بالبيت', 'الحج', 'الصفا', 'الهدى', 'هدى', 'يحل', "خرجنا"]	Topik ini jelas berkaitan dengan ibadah haji dan umrah, sebagaimana terlihat dari kata "الصفا", "الحج", dan "المروة". Topik ini menampilkan hadis-hadis yang menjelaskan tata cara pelaksanaan manasik, hukum ihram, dan syariat perjalanan ke Baitullah.

4.1.7 Evaluasi Model

Pada tahap evaluasi model ini interpretasi hasil di tampilkan dalam bentuk grafik seperti pada Gambar 4.8 di bawah ini yang didapatkan dari proses visulisai model hasil embeding integrasi penambahan fitur semantik berupa panjang teks dan TF-IDF.

Pada Gambar 4.8 menggambarkan bahwa setiap kluster terdiri atas hadis yang memiliki kesamaan konteks dan makna, yang dihasilkan dari analisis semantik yang lebih mendalam. Hasil ini sangat menarik mengingat BERT mampu menangkap kompleksitas bahasa yang digunakan dalam teks hadis, sehingga kelompok kluster memiliki koherensi yang tinggi berdasarkan studi yang relevan.



Gambar 4.8. Topik Word Scores

Gambar Topic Word Scores di atas diperjelas dengan keterangan legenda warna pada Gambar 4.9 Legenda Topik Word Scores menampilkan distribusi kata kunci utama pada masing-masing topik hasil klasterisasi hadis. Setiap topik direpresentasikan dengan warna berbeda agar mudah dibedakan. Warna oranye mewakili Topic 0 dan Topic 3, di mana kata-kata yang sering muncul antara lain ملكت، الإيمان، مع، قال، يا serta عن، حدثنا، قال، أبي، ابن, yang mengindikasikan hadis-hadis tentang sanad periwayatan dan pembahasan keimanan. Warna biru digunakan untuk Topic 1 dan Topic 4, dengan kata kunci seperti حديث، حسن، صحيح، هذا، مسلم pada Topic 1, serta ابيه، اخبرنا، خير، ابن، رؤية pada Topic 4, yang memperlihatkan adanya tema mengenai kualitas hadis dan riwayat perawi. Selanjutnya, warna ungu muda mewakili Topic 2 dengan kata kunci رحمة، المراة، شديدة، أبي، حدثنا yang menggambarkan hadis terkait kasih sayang dan perempuan. Warna hijau digunakan untuk Topic 5 dengan kata kunci أبي، يحيى، زكاة، عبد، سورة, yang menunjukkan adanya tema tentang zakat dan kewajiban ibadah. Warna kuning menunjukkan Topic 6, dengan kata-kata

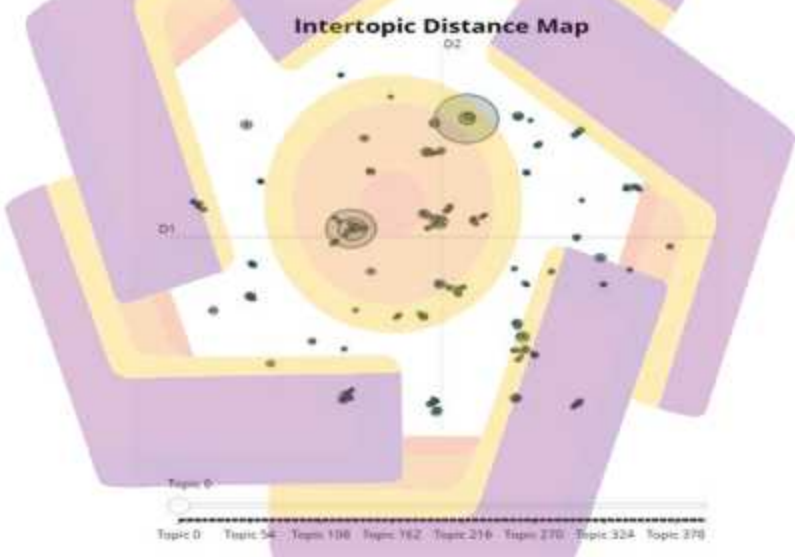
seperti قال، هذا، يا، صلى، الله yang sering ditemukan pada hadis berisi ucapan langsung Nabi Muhammad ﷺ. Terakhir, warna merah bata atau ungu tua digunakan untuk Topic 7, yang menampilkan kata-kata الرسول، رويا، الرجل، المؤمن، خير, sehingga dapat diasosiasikan dengan hadis mengenai mimpi, iman, dan kebaikan. Dengan adanya pembeda warna ini, pola kata kunci pada tiap topik menjadi lebih jelas, sekaligus membantu menginterpretasikan tema hadis yang terkandung dalam masing-masing kelompok.

Warna	Nomor Topik	Kata Kunci Utama
Oranye	Topic 0	عن، حدثنا، قال، أبي، ابن
Biru	Topic 1	حديث، حسن، صحيح، هذا، مسيح
Ungu Muda	Topic 2	رحمة، المرأة، شديدة، أبي، حدثنا
Oranye	Topic 3	ملكت، الإيمان، مع، قال، يا
Biru	Topic 4	أبيه، أخبرنا، خير، ابن، رويبة
Hijau	Topic 5	أبي، يحيى، زكاة، عبد، مسورة
Kuning	Topic 6	قال، هذا، يا، صلى، الله
Merah Bata / Ungu Tua	Topic 7	الرسول، رويا، الرجل، المؤمن، خير

Gambar 4.9. Legenda Topik Word Scores

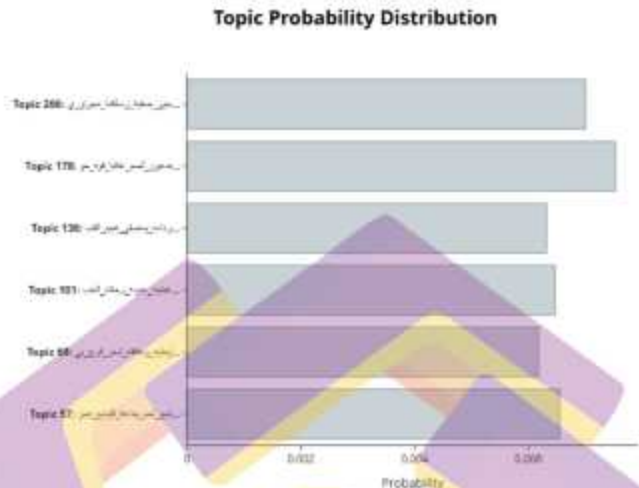
Gambar 4.10 Intertopic Distance Map menunjukkan peta jarak antar topik hasil klasterisasi hadis. Visualisasi ini umumnya dihasilkan dari reduksi dimensi menggunakan metode seperti t-SNE atau UMAP, yang memproyeksikan hubungan antar topik ke dalam bidang dua dimensi (D1 dan D2). Setiap lingkaran merepresentasikan satu topik, dengan ukuran lingkaran mencerminkan jumlah dokumen (hadis) yang termasuk dalam topik tersebut. Lingkaran yang lebih besar menandakan topik dengan jumlah hadis lebih banyak dan memiliki dominasi dalam korpus data. Sementara itu, jarak antar lingkaran menunjukkan tingkat kemiripan

antar topik. Topik yang berdekatan berarti memiliki kesamaan tema atau kosa kata, sedangkan topik yang berjauhan menandakan perbedaan yang signifikan dalam konten. Pada Gambar terlihat beberapa kelompok topik yang saling berdekatan, menandakan adanya keterkaitan erat, serta beberapa topik yang menyebar secara terpisah, menunjukkan tema khusus yang berbeda dari topik lainnya. Dengan demikian, peta ini membantu memberikan Gambaran menyeluruh mengenai distribusi, ukuran, serta keterkaitan antar topik hadis yang dihasilkan dari proses klasterisasi.



Gambar 4.10. Intropic Distance Map

Gambar 4.11 di baawah menunjukkan distribusi probabilitas topik dalam sebuah dokumen berdasarkan hasil analisis topic modeling. Dari grafik tersebut, dapat diketahui bahwa dokumen ini membahas beberapa topik sekaligus, namun dengan tingkat keterlibatan yang berbeda-beda.



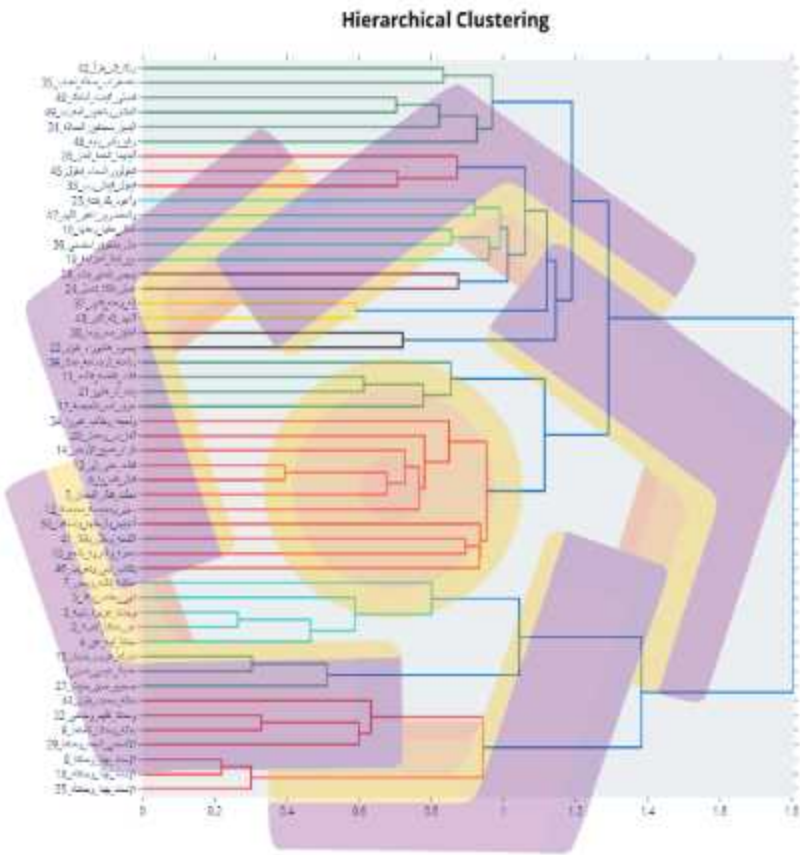
Gambar 4.11. Topik Probability Distribution

Topik yang paling dominan adalah Topic 178, probabilitas tertinggi (~0.0075), yang berarti topik ini memiliki kontribusi terbesar dalam isi dokumen. Ini menunjukkan bahwa sebagian besar isi dokumen berkaitan erat dengan kata kunci atau tema yang mewakili Topic 178.

Sementara itu, topik-topik lainnya, seperti Topic 68, 57, 111 dan lain-lain, juga muncul dalam dokumen, namun dengan probabilitas yang jauh lebih kecil. Hal ini menunjukkan bahwa meskipun dokumen menyentuh beberapa tema lain, namun topik-topik tersebut hanya dibahas secara terbatas atau tidak menjadi fokus utama.

Gambar 4.12. Hierarchical Clustering menunjukkan hasil pengelompokan data teks menggunakan metode hierarki klaster (Hierarchical Clustering). Teknik ini digunakan untuk mengelompokkan dokumen berdasarkan tingkat kemiripan kontennya. Hasil visualisasi ditampilkan dalam bentuk dendrogram, yaitu diagram

pohon yang menggambarkan proses penggabungan antar dokumen atau kelompok dokumen secara bertahap berdasarkan kesamaan isi.



Gambar 4.12. Hierarichal Clustering

Pada sumbu vertikal, ditampilkan nama-nama dokumen atau teks dalam bahasa Arab, sedangkan sumbu horizontal menggambarkan jarak atau perbedaan antar dokumen (dissimilarity). Semakin pendek garis horizontal antara dua cabang, semakin mirip dokumen-dokumen tersebut.

Warna-warna berbeda pada cabang dendrogram menandai terbentuknya beberapa kelompok utama (cluster). Masing-masing cluster menunjukkan sekelompok dokumen yang memiliki kemiripan topik atau kata kunci yang tinggi. Dengan demikian, dendrogram ini memudahkan dalam melihat bagaimana dokumen-dokumen tersebut saling berhubungan dan dapat dikelompokkan ke dalam tema atau kategori tertentu.

Secara keseluruhan, Gambar ini memberikan Gambaran struktur tematik dalam kumpulan data teks dan menunjukkan bahwa dokumen-dokumen yang dianalisis dapat diklasifikasikan ke dalam beberapa kelompok yang lebih kecil berdasarkan kesamaan isi.

Warna	Keterangan
Biru	Cluster utama dengan cabang besar, mencakup beberapa subkelompok hadis.
Hijau	Cluster dengan tema yang relatif mirip, tetapi terpisah dari biru pada jarak tertentu.
Merah	Cluster yang berisi kelompok hadis yang lebih homogen, terlihat cukup rapat.
Hitam	Cluster kecil yang berdiri sendiri, menandakan tema yang cukup unik.
Kuning	Cluster kecil yang spesifik, berbeda dari kelompok besar lainnya.
Warna tambahan (seperti cyan/ungu muda)	Sub-cluster lain yang terbentuk karena pemotongan hierarki lebih dalam.

Gambar 4.13. Legenda Hierarichal Clustering

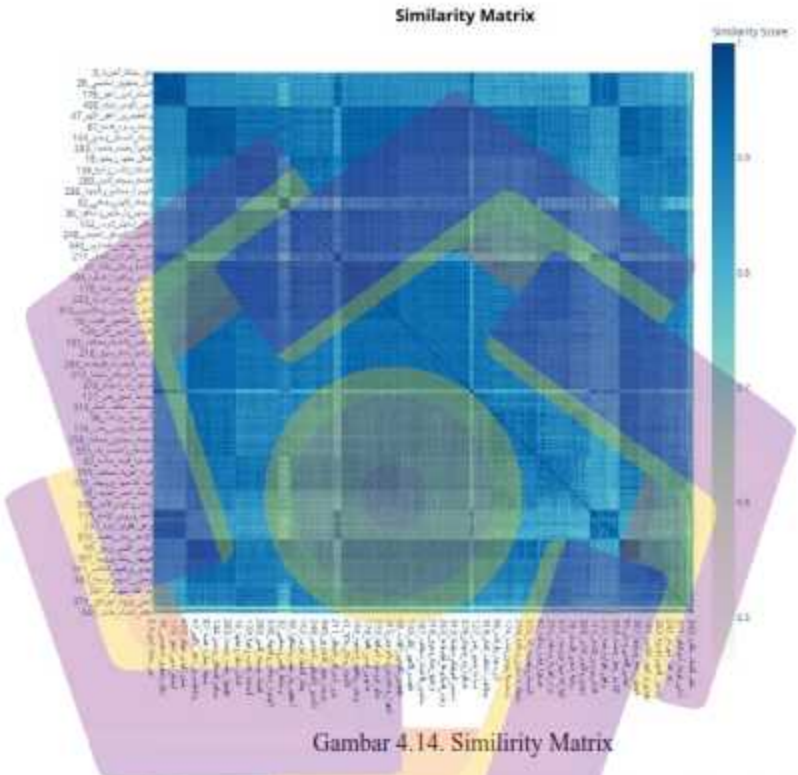
Hasil Hierarchical Clustering dengan keterangan Gambar 4.13 legenda Hierarical Clustering menunjukkan bahwa hadis-hadis dalam Shahih Bukhari dan Muslim dapat dikelompokkan ke dalam beberapa tema utama berdasarkan tingkat kemiripan kata kunci berdasarkan warna. Cluster dengan warna biru didominasi

oleh kata-kata seperti قال، حدثنا، عن، ابن yang merepresentasikan sanad periwayatan, sehingga kelompok ini dapat dipahami sebagai bagian dari struktur penyampaian hadis. Sementara itu, cluster hijau menampilkan kata kunci seperti رحمة، المرأة، شديدة yang menggambarkan tema akhlak dan hubungan sosial, khususnya terkait kasih sayang, peran perempuan, serta nilai moral. Cluster merah berfokus pada kata kunci seperti زكاة، عبد، مسورة yang erat kaitannya dengan ibadah dan kewajiban syariat, terutama zakat dan amalan sehari-hari. Adapun cluster kuning menonjolkan kata kunci الإيمان، الخير، الرجل، الرسول، رؤيا yang mengarah pada pembahasan iman dan keyakinan, termasuk hadis mengenai mimpi yang bernilai spiritual. Terakhir, cluster hitam terdiri dari kelompok kecil yang memiliki kosa kata khas dan berbeda dari mayoritas, sehingga dapat dipandang sebagai tema khusus seperti hadis hukum tertentu, kisah sahabat, atau topik unik yang jarang muncul. Dengan demikian, hasil klasterisasi hierarki ini memperlihatkan bahwa hadis tidak hanya terorganisir berdasarkan rantai periwayatan, tetapi juga dapat dipetakan ke dalam tema-tema substantif yang sesuai dengan isi kandungannya.

Pada Gambar 4.14. Similarity Matrix menampilkan matriks kemiripan antar dokumen teks dalam bentuk visualisasi heatmap. Matriks ini digunakan untuk menggambarkan sejauh mana tingkat kesamaan antara satu dokumen dengan dokumen lainnya dalam kumpulan data.

Setiap baris dan kolom pada matriks ini mewakili sebuah dokumen, dan warna pada titik pertemuan antara keduanya menunjukkan nilai skor kemiripan (similarity score). Warna yang digunakan berkisar dari biru tua hingga hijau muda, dengan biru tua menunjukkan tingkat kemiripan yang tinggi (nilai mendekati 1),

dan hijau muda menunjukkan tingkat kemiripan yang rendah (nilai mendekati 0.4 atau lebih rendah).



Gambar 4.14. Similirity Matrix

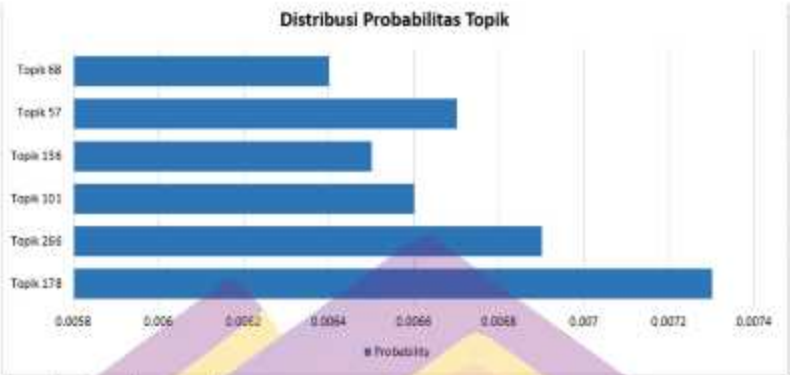
Terlihat jelas bahwa diagonal utama dari kiri atas ke kanan bawah selalu berwarna biru tua, karena menunjukkan perbandingan antara dokumen itu sendiri yang tentu memiliki kemiripan sempurna (nilai 1). Di luar diagonal, beberapa area memperlihatkan pola-pola blok berwarna lebih gelap, yang mengindikasikan adanya kelompok dokumen dengan kemiripan tinggi menunjukkan bahwa dokumen-dokumen tersebut mungkin membahas topik yang sama atau memiliki struktur isi yang serupa.

Berdasarkan pola warna yang muncul pada Similarity Matrix, terlihat adanya beberapa blok berwarna biru tua yang membentuk kotak-kotak tegas di dalam matriks. Blok-blok tersebut menunjukkan kumpulan hadis yang memiliki tingkat kesamaan sangat tinggi, sehingga dapat diinterpretasikan sebagai kelompok dengan tema yang sama. Misalnya, blok besar yang berada di bagian tengah kemungkinan merepresentasikan hadis tentang sanad periwayatan, karena kata-kata seperti ابن، عن، قال، حدثنا sering muncul secara konsisten dan membentuk kemiripan yang kuat antar dokumen. Sementara itu, blok lain yang lebih kecil tetapi cukup rapat dapat diasosiasikan dengan tema ibadah, seperti hadis tentang zakat atau doa, yang juga memiliki pola kosa kata khusus sehingga saling mengelompok. Di sisi lain, terdapat pula blok yang lebih terpisah dengan kesamaan tinggi di dalam kelompoknya, yang dapat dihubungkan dengan tema akhlak dan sosial, misalnya hadis tentang kasih sayang, perempuan, atau etika pergaulan. Selain itu, beberapa kotak kecil yang berada di pinggir matriks dengan warna biru pekat menandakan adanya cluster spesifik, misalnya hadis mengenai iman dan mimpi, yang meskipun jumlahnya tidak banyak, tetapi sangat homogen di dalam kelompoknya. Dengan demikian, Similarity Matrix ini tidak hanya menggambarkan tingkat kedekatan antar hadis secara numerik, tetapi juga memperlihatkan pemetaan tematik yang selaras dengan hasil klasterisasi sebelumnya, yaitu kelompok hadis tentang sanad, ibadah, akhlak, iman, dan tema-tema khusus.

4.2. Pembahasan

Proses validasi yang dilakukan pada metode terintegrasi fitur simantik menghasilkan metrik evaluasi silhouette score dengan nilai -0.1 dan Davies-Bouldin Index (DBI) 2.6, yang menunjukkan bahwa model klusterisasi yang dihasilkan memiliki kualitas yang laebih baik dibandingkan dengan tanpa terintegrasi yang menghasilkan silhouette score dengan nilai -0.145 dan nilai DBI 6.6. Sehingga skor pada metode terintegrasi cilhouette menunjukkan bahwa sebagian besar kluster memiliki separasi yang jelas dan tidak ada overlap yang signifikan antar kluster. Hasil ini membuktikan kehandalan metode yang digunakan, serta menunjukkan bahwa penambahan fitur semantik pada model BERT berkontribusi pada kedalaman hasil analisis. Nilai DBI sebesar 2.6 menunjukkan bahwa model klasterisasi bisa ditingkatkan, tetapi sudah menunjukkan peningkatan signifikan dibanding metode sebelumnya DBI = 6.6.

Interpretasi validasi hasil klusterisasi dapat dilihat pada Gambar 4.15 yang menunjukkan bahwa setiap kluster tidak hanya merepresentasikan sekelompok hadis, tetapi juga membentuk koneksi yang lebih besar dalam memahami ajaran Islam secara holistik. Pembaca atau peneliti dapat memperoleh wawasan baru tentang tema-tema yang saling berkaitan dalam tekstual hadis yang mungkin diabaikan dalam studi konvensional. Hal ini memperkuat fungsi hadis sebagai sumber pengetahuan yang kaya dan multifaset dalam konteks sosial dan keagamaan.



No.	Topic ID	Probability	Top Words (Ringkasan)	Kemungkinan Tema Topik
1	Topik 178	0.0073	عائشة, أبي, خرج, فجاء, ...	Kisah Aisyah & peristiwa mendadak
2	Topik 266	0.0069	صفة, شكل, جسم, عيسى, ...	Ciri-ciri fisik atau sifat Nabi / tokoh lainnya (mungkin Isa)
3	Topik 101	0.0066	عبيد, سعيد, طعيبة, رجال, ...	Sanad dan perawi melalui sahabat awal
4	Topik 136	0.0065	سمع, دعا, النبي, يسير, ...	Doa dan amalan saat bepergian
5	Topik 57	0.0067	حج, حجة, الأخمس, خرج, ...	Haji Wada' dan perjalanan terakhir Nabi
6	Topik 68	0.0064	ناقة, بعير, وابره, ...	Hewan tunggangan Nabi dalam perjalanan

Gambar 4.15. Distribusi Probabilitas Topik-Dokumen Hadis

Interpretasi Topik hasil klasterisasi dibandingkan dengan klasifikasi Bab pada kitab hadis Bukhori Muslim menunjukkan adanya keterkaitan topik. Jika dilihat pada kitab hadis Bukhori Muslim ke 147 seperti pada Gambar 4.16 mengenai Bab Thoharoh yang didalamnya menerangkan wudhu. Maka pada topik klatering topik 153 dengan hadis yang sama menerangkan :

[«لحاجته»، «صبي»، «إبريق»، «المغيرة»«بإبريق»، «حاجته»، «تبعه»، «مربع»، «ماء»، «إناء»]

["Untuk keperluannya", "anak laki-laki", "kendi", "al-Mughira", "dengan kendi", "keperluannya", "mengikutinya", "persegi", "air", "bejana"].

Gambar 4.16. Hadis No. 147 tentang Thoharoh

Dalam merepresentasikan Topik Hadis dengan Klasterisasi topik di representasikan dalam bentuk kata bukan kalimat seperti pengelompokan Bab maupun Sub Bab yang ada di Kitab Hadis Bukhori Muslim seperti tampak pada Gambar 4.17.

Gambar 4.17. Klasterisasi Hadis No. 151 topik 153

Pada Gambar 4.17 di atas menunjukkan bahawa dalam kecocokan topik memiliki kemiripannya dalam hal air yaitu jika pada kitab Bukhori dikelompokkan

dalam Bab Wudlu tetapi jika diklasterisasi masuk topik air dengan kesamaan dalam berwudhlu menggunakan air.

Sehingga terkait dalam penelitian ini metode klasterisasi belum bisa menyimpulkan ke dalam bentuk topik seperti yang ada di Kitab Bukhori Muslim yang bisa merepresentasikan ke dalam bentuk Bab Wudlu.

BAB V

PENUTUP

5.1. Kesimpulan

Penelitaian pengembangan model klusterisasi topik hadis Bukhari Muslim menggunakan BERT dengan penambahan fitur semantik menunjukkan bahwa pendekatan ini dapat diterapkan secara efektif terbukti dengan munculnya beberapa topik yang dominan seperti formula pembuka periwayatan dan pengenalan Rasulullah dalam sanad. Sedangkan akurasi dan interpretabilitas dalam mengelompokkan hadis berdasarkan tema dan konteksnya dibuktikan menggunakan metrik evaluasi silhouette score dengan nilai -0.1 dan davies-bouldin index (DBI) 2.6. dan melalui tahapan metodologi yang telah dilaksanakan

5.2. Saran

Idealnya, agar klaster benar-benar representatif dan terpisah dengan baik Nilai DBI perlu diturunkan lebih lanjut, mendekati < 1.5, dan meningkatkan nilai silhouette score agar lebih mendekati 1. Dan secara keseluruhan, penelitian ini tidak hanya memberikan kerangka kerja baru untuk klusterisasi topik hadis menggunakan teknologi canggih, tetapi juga membuka peluang bagi penelitian lebih lanjut dalam bidang ini, dengan mempertimbangkan integrasi antara teologi dan implementasi teknologi dalam studi agama. Temuan-temuan yang diperoleh dapat digunakan sebagai dasar untuk mengembangkan perangkat pembelajaran berbasis AI dalam pengajaran hadis dan pemahaman ajaran Islam di kalangan generasi muda.

DAFTAR PUSTAKA

[1] H. M. Abdelaal, B. R. Elemary, and H. Youness, "Classification of Hadith According to Its Content Based on Supervised Learning Algorithms," *Ieee Access*, vol. 7, pp. 152379–152387, 2019, doi: 10.1109/access.2019.2948159.

[2] S. E. Pratama, W. Darmalaksana, D. S. Maylawati, H. Sugilar, T. Mantoro, and M. A. Ramdhani, "Weighted Inverse Document Frequency and Vector Space Model for Hadith Search Engine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, p. 1004, 2020, doi: 10.11591/ijeecs.v18.i2.pp1004-1014.

[3] A. S. Azmi, "The Influence of Abbasid Empire and Community Needs in the Development of Hadīth Literature and Islamic Prophetology," *Umran - International Journal of Islamic and Civilizational Studies*, vol. 4, no. 2, 2017, doi: 10.11113/umran2017.4n2.138.

[4] A. H. Usman, R. Wazir, and Z. Ismail, "The Notion of Liberalisation on the Anti-Hadith Movement and Its Impact on Society," *Al-Irsyad Journal of Islamic and Contemporary Issues*, vol. 2, no. 2, pp. 81–94, 2017, doi: 10.53840/alirsyad.v2i2.20.

[5] A. Duderija, "Evolution in the Canonical Sunni Hadith Body of Literature and the Concept of an Authentic Hadith During the Formative Period of Islamic Ought as Based on Recent Western Scholarship," *Arab Law Quarterly*, vol. 23, no. 4, pp. 389–415, 2009, doi: 10.1163/157302509x467371.

[6] N. Nikmatullah, "Male Ulama Reinterpretation of the Gender Hadith in Indonesian Socio Cultural Contexts," *Pharos Journal of Theology*, no. 105(2), 2024, doi: 10.46222/pharosjot.105.213.

[7] N. S. Hamisan@Khair, "The Role of Women and the Contextualization of Peace in Modern Times: Analysis on Hadith Perspective," *Umran - International Journal of Islamic and Civilizational Studies*, vol. 10, no. 3, pp. 57–72, 2023, doi: 10.11113/umran2023.10n3.617.

[8] T. Tasbih, "Islamic Feminists' Rejection of the Textual Understanding of Misogynistic Hadiths for the Advancement of Gender Justice in Makassar, Indonesia," *Samarah Jurnal Hukum Keluarga Dan Hukum Islam*, vol. 8, no. 1, p. 196, 2024, doi: 10.22373/sjhk.v8i1.19856.

[9] F. Binbeshr, A. Kamsin, and M. Mohammed, "A Systematic Review on Hadith Authentication and Classification Methods," *Acm Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 2, pp. 1–17, 2021, doi: 10.1145/3434236.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[11] H. Wang, L. Xu, and X. He, "Topic modeling in the era of large language models: A survey," *Information Fusion*, vol. 94, pp. 101–120, 2023, doi: 10.1016/j.inffus.2023.101904.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings*

of NAACL-HLT 2019, 2019, pp. 4171–4186. doi: 10.48550/arXiv.1810.04805.

[13] L. George, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology*, vol. 15, pp. 1121–1130, 2023, doi: 10.1007/s41870-023-01599-x.

[14] M. Grootendorst, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. doi: 10.48550/arXiv.2203.05794.

[15] Y. Ortakci, A. Cetinkaya, and H. Ozkose, "Investigating transfer learning capacity of SBERT models through pooling for text clustering," *Journal of Information Security and Applications*, vol. 76, p. 103558, 2024, doi: 10.1016/j.jisa.2024.103558.

[16] M. Hankar, "A comprehensive overview of topic modeling," *Neurocomputing*, 2025, doi: 10.1016/j.neucom.2025.127654.

[17] A. Suryawan, "Challenges in supervised and unsupervised learning," *Int J Adv Sci Eng Inf Technol*, vol. 14, no. 1, pp. 55–66, 2024, doi: 10.18517/ijaseit.14.1.20723.

[18] A. M. Farooqi, "Multi-IsnadSet (MIS) for Sahih Muslim Hadith with chain of narrators," *Data Brief*, 2024, doi: 10.1016/j.dib.2024.110123.

[19] S. Aftar, A. Elmadany, and T. Elsayed, "A novel approach for topic extraction in Islamic studies," *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, doi: 10.18653/v1/2024.findings-emnlp.###.

[20] A. M. Ikotun, "K-Means clustering algorithms: A comprehensive review and applications," *Inf Sci (N Y)*, vol. 627, pp. 128–152, 2023, doi: 10.1016/j.ins.2023.04.015.

[21] Y. Ortakci, A. Cetinkaya, and H. Ozkose, "Optimizing SBERT for long text clustering: Two novel approaches," *J Supercomput*, 2025, doi: 10.1007/s11227-025-06530-3.

[22] R. Dodda, "BERT-based Document Clustering: Unveiling Semantic Patterns in 20News Group, Reuters, and BBC Sports Corpora," 2024, doi: 10.22541/au.171506422.20645846/v1.

[23] A. Gupta and V. Gupta, "Unsupervised Contextualized Document Representation," pp. 166–173, 2021, doi: 10.18653/v1/2021.sustainlp-1.17.

[24] A. Subakti, H. Murfi, and N. Hariadi, "The Performance of BERT as Data Representation of Text Clustering," *J Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00564-9.

[25] L. George and P. Sumathy, "An Integrated Clustering and BERT Framework for Improved Topic Modeling," 2022, doi: 10.21203/rs.3.rs-1986180/v1.

[26] A. K. Yadav, T. Gupta, M. Kumar, and D. Yadav, "A Hybrid Model Integrating LDA, BERT, and Clustering for Enhanced Topic Modeling," *Qual Quant*, 2025, doi: 10.1007/s11135-025-02077-y.

[27] N. Peinelt, D. Nguyen, and M. Liakata, "tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection," 2020, doi: 10.18653/v1/2020.acl-main.630.

[28] T. Sun, X. Liu, X. Qiu, and X. Huang, "Paradigm Shift in Natural Language Processing," *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022, doi: 10.1007/s11633-022-1331-6.

[29] N. Z. binti M. Ismail, "Systematic Literature Review on the Ontology Representation Related the Hadith Corpus," *Journal of Hadith Studies*, pp. 105–116, 2023, doi: 10.33102/johs.v8i2.250.

[30] M. Muhammed, "Arabic Ontology for Hadith Texts - A Survey," *The Egyptian Journal of Language Engineering*, vol. 11, no. 1, pp. 1–14, 2024, doi: 10.21608/ejle.2024.266774.1062.

[31] M. M. A. Najeeb, "Towards a Deep Leaning-Based Approach for Hadith Classification," *European Journal of Engineering and Technology Research*, vol. 6, no. 3, pp. 9–15, 2021, doi: 10.24018/ejeng.2021.6.3.2378.

[32] T. Liu, H. Yu, and R. H. Blair, "Stability Estimation for Unsupervised Clustering: A Review," *Wiley Interdiscip Rev Comput Stat*, vol. 14, no. 6, 2022, doi: 10.1002/wics.1575.

[33] E. R. Watson, A. Mora, A. T. Fard, and J. C. Mar, "How Does the Structure of Data Impact Cell–cell Similarity? Evaluating How Structural Properties Influence the Performance of Proximity Metrics in Single Cell RNA-seq Data," *Brief Bioinform*, vol. 23, no. 6, 2022, doi: 10.1093/bib/bbac387.

[34] L. George and P. Sumathy, "An Integrated Clustering and BERT Framework for Improved Topic Modeling," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, 2023, doi: 10.1007/s41870-023-01268-w.

[35] J. Devlin, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/arxiv.1810.04805.

[36] A. Raganato and J. Tiedemann, "An Analysis of Encoder Representations in Transformer-Based Machine Translation," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupała, and A. Alishahi, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 287–297. doi: 10.18653/v1/W18-5431.

[37] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A Joint Model for Video and Language Representation Learning," 2019, doi: 10.1109/iccv.2019.00756.

[38] S. Smith *et al.*, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, a Large-Scale Generative Language Model," 2022, doi: 10.48550/arxiv.2201.11990.

[39] G. Emerson, "Autoencoding Pixies: Amortised Variational Inference With Graph Convolutions for Functional Distributional Semantics," 2020, doi: 10.18653/v1/2020.acl-main.367.

[40] S. Serrano and N. A. Smith, "Is Attention Interpretable?," 2019, doi: 10.18653/v1/p19-1282.

[41] M. Choi *et al.*, "MelBERT: Metaphor Detection via Contextualized Late Interaction Using Metaphorical Identification Theories," 2021, doi: 10.18653/v1/2021.naacl-main.141.

[42] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," 2019, doi: 10.18653/v1/d19-1077.

[43] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," 2021, doi: 10.48550/arxiv.2108.07258.

[44] Y. Ji, Z. Zhou, H. Liu, and R. V Davuluri, "DNABERT: Pre-Trained Bidirectional Encoder Representations From Transformers Model for DNA-language in Genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021, doi: 10.1093/bioinformatics/btab083.

[45] J. D. Moffitt, C. King, and K. M. Carley, "Hunting Conspiracy Theories During the COVID-19 Pandemic," *Soc Media Soc*, vol. 7, no. 3, p. 205630512110432, 2021, doi: 10.1177/20563051211043212.

[46] H. Guo, T. Huang, H. Huang, M. Fan, and G. Friedland, "Detecting COVID-19 Conspiracy Theories With Transformers and TF-IDF," 2022, doi: 10.48550/arxiv.2205.00377.

[47] L. Tilton and T. Arnold, "An Introduction Natural Language Processing," *RMC*, vol. 1, no. 1, 2019, doi: 10.46632/rmc/1/1/014.

[48] J. Hirschberg and C. D. Manning, "Advances in Natural Language Processing," *Science (1979)*, vol. 349, no. 6245, pp. 261–266, 2015, doi: 10.1126/science.aaa8685.

[49] M. S. Keezhatta, "Understanding EFL Linguistic Models Through Relationship Between Natural Language Processing and Artificial Intelligence Applications," *Arab World English Journal*, vol. 10, no. 4, pp. 251–262, 2019, doi: 10.24093/awej/vol10no4.19.

[50] Z. Jin, G. Chauhan, B. Tse, M. Sachan, and R. Mihalcea, "How Good Is NLP? A Sober Look at NLP Tasks Through the Lens of Social Impact," 2021, doi: 10.48550/arxiv.2106.02359.

[51] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and Efficient Fine-Tuning for Pre-Trained Natural Language Models Through Principled Regularized Optimization," 2020, doi: 10.18653/v1/2020.acl-main.197.

[52] N. Chomsky, "The Minimalist Program," 2014, doi: 10.7551/mitpress/9780262527347.001.0001.

[53] A. W. Hidayat, M. Z. M. Mustaqim, B. A. Hadi, and A. Hashim, "The Influence of Al-Sahihayn on Popular Hadith Literatures: The Case of Khazinah Al-AsrarJalilah Al-Adhkar," *Global Journal Al-Thaqafah*, vol. 7, no. 1, pp. 29–37, 2017, doi: 10.7187/gjat12420170701.

[54] E. H. Muchtar, "Analisis Deskriptif Kitab Shahih Al-Bukhari," *Jiqta Jurnal Ilmu Al-Qur an Dan Tafsir*, vol. 1, no. 1, pp. 19–34, 2022, doi: 10.36769/jiqta.v1i1.187.

[55] S. M. M. Zin, "في صناعة الأسانيد وتحويلها للمتن الواحد (البخاري ومسلم)براعة الشيخان," *Jurnal Yadim*, vol. 2, no. 1, 2023, doi: 10.61465/jurnalyadim.v2.82.

[56] G. Murtaza and F. N. Alvi, "معنعن روایت کے بارے میں امام بخاری اور امام مسلم کا منہج: صحیحین کے تناظر میں," *Al Basirah*, vol. 11, no. 01, pp. 67–90, 2022, doi: 10.52015/albasirah.v11i01.9.

[57] S. R. M. Najib, N. A. Rahman, N. K. Ismail, N. Alias, Z. M. Nor, and M. N. Alias, "Comparative Study of Machine Learning Approach on Malay

Translated Hadith Text Classification Based on Sanad," *Matec Web of Conferences*, vol. 135, p. 00066, 2017, doi: 10.1051/matecconf/201713500066.

[58] M. A. Z. Yaakob *et al.*, "An Analysis of Waqf Hadiths in Sahih Al-Bukhari Per Fiqh Al-Bukhari Perspective," *International Journal of Academic Research in Business and Social Sciences*, vol. 12, no. 11, 2022, doi: 10.6007/ijarbss/v12-i11/15416.

[59] T. Hidayat and E. Sumarna, "Hadis Kehujjahan Hadis Menurut Imam Empat MaÅ¼hab (Studi Analisa Terhadap Metode Penyusunan Al-Kutub Al-SittaÄ§)," *Religia*, pp. 115–135, 2019, doi: 10.28918/religia.v22i1.1386.

[60] B. Sumintono, E. S. Kusumaputri, H. Hariri, and Y. Juniardi, "Islamic Educational Leadership," pp. 159–175, 2023, doi: 10.4324/9781003360070-12.

[61] T. Alam and J. Schneider, "Social Network Analysis of Hadith Narrators From Sahih Bukhari," 2020, doi: 10.1109/besc51023.2020.9348299.

[62] S. N. Khan, "Modesty During Childbirth : Perspectives of Immigrant Muslim Women in Canada," 2021, doi: 10.32920/ryerson.14654250.v1.

[63] A. Mohtarom, "Struggling Against Radicalism Through the Sunnah of the Prophet Muhammad SAW.," *J. Multidisiplin Ibrahimy*, vol. 1, no. 1, pp. 14–25, 2023, doi: 10.35316/jummy.v1i1.3496.

[64] N. H. Pulungan, "An Orientalist Today: Jonathan A.C. Brown's Thoughts on Hadith," *Ulumuna*, vol. 27, no. 2, pp. 552–572, 2023, doi: 10.20414/ujis.v27i2.767.

[65] S. Mahmoud, O. Saif, E. Nabil, M. Abdeen, M. ElNainay, and M. Torki, "AR-Sanad 280K: A Novel 280K Artificial Sanads Dataset for Hadith Narrator Disambiguation," *Information*, vol. 13, no. 2, p. 55, 2022, doi: 10.3390/info13020055.

[66] E. Hafid and M. Mahmuddin, "Criticism of Hadith Authenticity on Contemporary Islamic Thinkers," *Journal of Islam and Science*, vol. 9, no. 2, pp. 119–126, 2023, doi: 10.24252/jis.v9i2.31696.

[67] S. Ashar, "Implementasi Penilaian Otentik Dalam Pembelajaran Al Qur'an Hadits Di MTs Salafiyah Bidayatul Hidayah Mojogeneng Jatirejo Mojokerto," *Progressa Journal of Islamic Religious Instruction*, vol. 1, no. 2, p. 7, 2018, doi: 10.32616/pgr.v1.2.71.7-14.

[68] A. K. Bangash and W. Ali, "Perceptions of Islamic Scholars Towards Women's Political Empowerment in Pashtun Society (A Case Study of District Mardan)," *Journal of Religious Studies (Uochjrs)*, pp. 37–56, 2018, doi: 10.33195/uochjrs-v2iiii962018.

[69] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1201.0490

[70] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[71] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge: Cambridge University Press, 2011. doi: DOI: 10.1017/CBO9781139058452.

[72] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. doi: DOI: 10.1017/CBO9780511809071.

[73] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," Apr. 13, 2016, *Royal Society of London*. doi: 10.1098/rsta.2015.0202.

[74] J. Shlens, "A Tutorial on Principal Component Analysis," Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.1100

[75] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," 1987.

[76] J. Han, M. Kamber, and J. Pei, "Preface," in *Data Mining: Concepts and Techniques (Third Edition)*, J. Han, M. Kamber, and J. Pei, Eds., Boston: Morgan Kaufmann, 2012, pp. xxiii–xxix. doi: https://doi.org/10.1016/B978-0-12-381479-1.00020-4.

[77] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans Pattern Anal Mach Intell*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[78] P.-Nin. Tan, Michael. Steinbach, and Vipin. Kumar, *Introduction to data mining*. Pearson, 2018.