Vol. 4 No. 2 – Tahun 2025
Indonesian Journal Computer Science
ISSN: 2829-3819

Pengembangan Model Klasterisasi Topik Hadis Bukhari Muslim Menggunakan BERT dengan Penambahan Fitur Semantik

Ahmad Hasyim Asy'ari[1], Muhammad Hanafi[2]
Universitas Amikom Yogyakarta[1,2]
ahmadhasyim@students.amikom.ac.id[1], hanafi@amikom.ac.id[2]

Diterima	Direvisi	Disetujui
(20-05-2025)	(31-05-2025)	(03-06-2025)

Abstrak - 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.*

Keywords: Hadis, Semantic Features, BERTopic, NLP

I. PENDAHULUAN

Pengembangan model klusterisasi topik hadis Bukhari Muslim merupakan langkah inovatif dalam upaya menyusun dan memahami ajaran Islam dengan lebih sistematis dan terstruktur. Hadis sebagai sumber kedua dalam ajaran Islam memberikan pedoman penting bagi masyarakat Muslim, baik dalam aspek spiritual maupun sosial. Oleh karena itu, analisis mendalam terhadap koleksi hadis ini, baik dari segi konten maupun konteks, sangat penting untuk meningkatkan pemahaman serta penerapan nilai-nilai dalam kehidupan sehari-hari. Dalam konteks kajian hadis, pengklasteran topik menggunakan model pemrosesan bahasa alami, seperti BERT *(Bidirectional Encoder Representations from Transformers),* telah menjadi fokus penelitian yang semakin penting. Hadis, yang merupakan ucapan dan tindakan Nabi Muhammad SAW, memiliki berbagai tema

dan konteks yang perlu dianalisis secara mendalam untuk menghasilkan pemahaman yang lebih baik (Mudding, 2024). Dengan kemampuannya dalam menghasilkan penyematan kontekstual yang kaya, BERT menawarkan potensi besar untuk mengklaster informasi ini dengan lebih efektif dibandingkan metode tradisional (Maulida, 2023).

Model BERT telah muncul sebagai salah satu inovasi utama dalam pengolahan bahasa alami *(Natural Language Processing - NLP),* memberikan kemampuan untuk menghasilkan penyematan yang kontekstual dan representatif dari teks. Kemampuan BERT untuk memahami konteks kata dalam kalimat, serta merepresentasikan hubungan antar kata, menjadikannya sangat efektif untuk berbagai tugas NLP seperti pengklasteran dokumen dan pemodelan topik (Asy'ari et al., n.d.); (Murfi et al., 2024); (Subakti et al., 2022). Model ini memanfaatkan mekanisme perhatian (attention mechanism) yang memungkinkan model untuk memproses setiap kata dalam konteks kata-kata sekitarnya, sehingga meningkatkan akurasi dalam memahami makna yang lebih dalam dari teks (Yang, 2024); (Subakti et al., 2022).

Namun, satu tantangan yang dihadapi dalam menggunakan BERT adalah ukurannya yang besar dan kompleksitas komputasi yang diperlukan, yang dapat menjadi kendala dalam penerapan pada tugas dengan sumber daya terbatas. Berbagai penelitian telah menunjukkan bahwa meskipun BERT menawarkan penyematan kontekstual yang kaya, penggunaannya secara eksklusif berpotensi mengabaikan fitur linguistik penting lainnya, yang dapat mempengaruhi kinerja hasil pengelompokan (Dodda & Alladi, 2024);, (George & Sumathy, 2023).

Penelitian terdahulu Model Klasterisasi Topik Hadis Bukhari Muslim Menggunakan BERT sudah dilakukan oleh (Asy'ari et al., n.d.) namun di penelitiannya masih belum dilakukan analisa secara mendalam dengan mencantumlan nilai *silhouette score* dan *davies-bouldin Index*. Di penelitian tersebut hanya memvisualisasikan hasil klastering saja dan memfokuskan pada keberhasilannya dalam menerapkan klasterisasi berdasarkan topik yang dominan. Peneliti menyarankan untuk meningkatkan BERT dengan berbagai cara untuk meningkatkan model klasterisasinya.

Misalnya, beberapa penelitian menyarankan bahwa meningkatkan BERT dengan memasukkan fitur tambahan seperti embedding entitas dapat lebih meningkatkan efektivitas dalam tugas tertentu yang berorientasi pada entitas (Gerritse, 2022).

Sebagai respons terhadap kendala tersebut, beberapa studi telah menyarankan pengembangan BERT dengan teknik lain seperti penambahan fitur semantik yang relevan dapat menyebabkan informasi penting yang berkaitan dengan bahasa dan konteks hadis tidak terabaiakan sehingga meningkatkan efektivitas peningkatan pemahaman kontekstual (Rohman, 2021).

Penambahan fitur semantik dalam model klasterisasi hadis bukan hanya meningkatkan pemahaman kontekstual, tetapi juga memperkuat kemampuan model dalam menyampaikan nuansa dan makna yang terkandung dalam hadis. Dalam pendekatan ini, fitur semantik seperti hubungan antar entitas, asumsi konteks, dan pola bahasa dapat ditambahkan untuk memperbaiki representasi informasi yang dihasilkan oleh model (Rinjani, 2021). Dengan demikian dapat membuat pengelompokan yang lebih bermakna, yang tidak hanya berdasarkan kata kunci atau frasa, tetapi juga memperhatikan hubungan semantik antara hadis-hadis yang berbeda (Imana et al., 2024).

Beberapa studi mengindikasikan bahwa tanpa integrasi fitur semantik, model pengklusteran mungkin mengalami keterbatasan dalam menciptakan kelompok yang koheren dan relevan. Misalnya, hadis yang berbagi tema moral dan etika mungkin tidak terkelompok dengan baik jika hanya mempertimbangkan struktur bahasa dasar dari teks tanpa memperhatikan makna mendalam yang tersirat di dalamnya (Yulianingsih & Nursihah, 2021). Selain itu, analisis semantik yang lebih mendalam juga dapat memberikan wawasan baru yang relevan dalam kajian hadis yang dapat membantu dalam interpretasi dan penerapan hadis dalam konteks modern (Riantika, 2023).

Dalam upaya untuk meningkatkan efektivitas pengklasteran topik hadis Bukhari dan Muslim, penggunaan fitur semantik tambahan seperti panjang teks, *Term Frequency* (TF), dan I*nverse Document Frequency* (IDF) menjadi sangat penting. Model BERT sudah terbukti ampuh dalam menangkap konteks linguistik yang kompleks, dan penambahan fitur-fitur semantik tersebut dapat meningkatkan kualitas hasil klasterisasi. Setiap fitur berfungsi untuk meningkatkan representasi semantik teks dan memberikan pemahaman yang lebih mendalam terhadap hubungan antar hadis.

Panjang Teks merupakan fitur yang menunjukkan seberapa banyak informasi yang terkandung dalam setiap hadis. Dalam konteks klasterisasi, panjang teks dapat memengaruhi cara model BERT memahami konteks dan topik. Beberapa penelitian menunjukkan bahwa segmen teks yang lebih panjang memiliki kecenderungan untuk mengandung lebih banyak informasi kontekstual yang relevan (Subakti et al., 2022). Fitur ini dapat membantu dalam membedakan antara hadis-hadis yang

lebih kompleks dari yang lainnya, sehingga dapat meningkatkan pemisahan antar kelompok dalam pengklasteran.

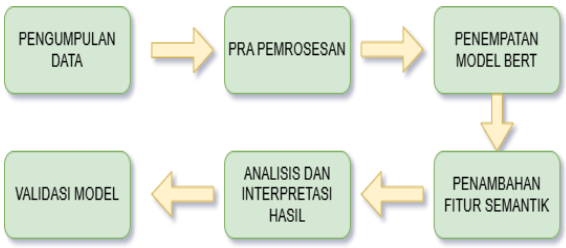
Term Frequency (TF) adalah ukuran yang menyoroti seberapa sering suatu istilah muncul dalam sebuah dokumen. Dalam pengklasteran topik, TF berfungsi untuk memberikan bobot lebih pada kata-kata yang sering muncul dalam setiap hadis. Penelitian menunjukkan bahwa penggunaan TF dalam kombinasi dengan BERT dapat membantu dalam meningkatkan akurasi pengelompokan dengan memberikan penekanan pada kata kunci yang paling relevan dalam konteks topik tertentu (Subakti et al., 2022). Hal ini meningkatkan daya ungkit model BERT dalam pengelompokan yang lebih efektif.

Inverse Document Frequency (IDF) berfungsi untuk mengurangi bobot kata-kata yang sering muncul di banyak dokumen, tetapi jarang muncul di dokumen tertentu. IDF membantu dalam menonjolkan kata-kata yang lebih unik dan khusus untuk konteks tertentu, sehingga lebih penting dalam proses pengklasteran yang berfokus pada hadis yang memiliki makna mendalam (Zhou et al., 2021). Dengan mempertimbangkan IDF, pengelompokan hadis dapat dioptimalkan untuk memastikan bahwa kelompok yang dibentuk lebih informatif dan relevan.

Sebuah studi menunjukkan bahwa penggabungan fitur semantik seperti TF dan IDF dengan BERT dapat memperbaiki kemampuan model dalam menangkap pola semantik yang kaya, yang pada akhirnya menghasilkan pengelompokan yang lebih baik dengan tingkat kohesi yang lebih tinggi antara dokumen dalam grup yang sama (Zhou et al., 2021). Oleh karena itu, integrasi fitur-fitur semantik ini dalam proses klasterisasi topik hadis Bukhari dan Muslim menggunakan BERT diharapkan dapat menghasilkan analisis yang lebih mendalam dan efisien serta memberikan wawasan yang lebih akurat dalam memahami hadis-hadis tersebut.

II. METODOLOGI PENELITIAN

Metodologi penelitian dalam pengembangan model klusterisasi topik hadis Bukhari Muslim menggunakan BERT dengan penambahan fitur semantik terdiri dari enam tahapan utama yang saling berhubungan.



Sumber : Hasil Penelitaian (2025)

Gambar 1. Tahapan Penelitian

Pada setiap segmen pada metodologi ini di buat didasarkan pada kebutuhan untuk memastikan bahwa setiap langkah dalam proses analisis data dilakukan dengan cermat dan sistematis.

1. Pengumpulan Data:

Langkah pertama adalah mengumpulkan data berupa dataset dari https://www.kaggle.com/ dan bersifat open source, dimana data hadis tersebut berasal dari karya-karya Bukhari dan Muslim. Proses ini melibatkan pengumpulan teks lengkap dan metadata yang relevan. Sumber-sumber harus diverifikasi untuk keakuratan dan keterandalan, sehingga menciptakan fondasi yang solid untuk analisis selanjutnya. Penggunaan dataset yang kaya akan variasi konten akan mempengaruhi hasil klusterisasi yang dihasilkan (Liu et al., 2022).

2. Pra-pemrosesan:

Setelah pengumpulan data, tahap selanjutnya mencakup pra-pemrosesan yang meliputi pembersihan teks, penghapusan karakter khusus, normalisasi teks, tokenisasi, dan penghilangan stopwords. Teknik lemmatization juga akan diterapkan untuk menyatukan kata-kata yang memiliki bentuk berbeda namun memiliki makna yang sama. Proses ini diharapkan dapat meningkatkan kualitas input yang digunakan dalam model BERT sehingga hasil klusterisasi menjadi lebih relevan (Li et al., 2024).

3. Penerapan Model BERT:

Tahap berikutnya adalah penerapan model BERT untuk menghasilkan representasi semantik dari data hadis yang telah dipra-pemrosesan. BERT akan mengubah teks menjadi vektor numerik yang dapat merepresentasikan konteks kata dalam kalimat secara hirarkis. Ini penting karena model ini memiliki kemampuan untuk memahami makna kata dalam konteks yang lebih dalam, yang selanjutnya akan meningkatkan akurasi klusterisasi dibandingkan dengan metode tradisional (Hua, 2024).

4. Penambahan Fitur Semantik untuk Klusterisasi:

Setelah mendapatkan representasi dari BERT, fitur semantik tambahan akan diintegrasikan untuk memberikan konteks tambahan pada data. Ini mungkin meliputi penghitungan semantik berbasis konsep seperti persentase kemunculan topik dalam kumpulan data. Klusterisasi akan dilakukan dengan memanfaatkan algoritma seperti K-Means atau DBSCAN, yang dapat memanfaatkan fitur semantik ini untuk meningkatkan pemahaman

tentang hubungan antar hadis (Aluri & Latha, 2023). Pada bagian ini model akan ditambahkan fitu panjang teks dan TF-IDF pada data embeding agar bisa diproses ke tahap berikutnya.

5. Analisis dan Interpretasi Hasil:

Setelah proses klusterisasi selesai, langkah selanjutnya adalah analisis hasil kluster yang dihasilkan. Setiap kluster akan dievaluasi berdasarkan tema dan konten yang tersimpan di dalamnya. Analisis ini akan berfokus pada identifikasi pola-pola yang relevan dan pengelompokan hadis berdasarkan makna yang lebih mendalam, memberikan wawasan baru terkait hubungan antar ajaran dalam hadis tersebut. Hasil klusterisasi yang telah diinterpretasikan dapat memberikan perspektif yang berharga bagi pengkajian keilmuan (Aminah et al., 2025). Pada bagaian ini akan dibuat visualisasi berupa grafik model visualisasi topik hadis yang akan di representasikan pada Hasil dan Pembahasan.

6. Validasi Model:

Sebagai langkah akhir, model yang telah dikembangkan perlu divalidasi menggunakan metrik evaluasi seperti Silhouette score atau Davies-Bouldin index untuk menilai kualitas kluster yang terbentuk. Proses ini penting untuk memastikan bahwa model tidak hanya berfungsi secara teoritis, tetapi juga praktis dan bermanfaat dalam aplikasi dunia nyata (Shen et al., 2024). Validasi model pada tahap ini sampai pada penampilan hasil *Silhouette score* dan *davies-bouldin Index*.

Dengan menerapkan metodologi yang sistematis ini, penelitian diharapkan dapat menghasilkan model klusterisasi yang tidak hanya akurat tetapi juga bermanfaat dalam memberikan interpretasi yang lebih mendalam terhadap hadis Bukhari dan Muslim serta membantu dalam edukasi dan pemahaman ajaran Islam.

III. HASIL DAN PEMBAHASAN

Penelitaian pengembangan model klusterisasi topik hadis Bukhari Muslim menggunakan BERT dengan penambahan fitur semantik dibahas melalui beberapa tahapan penelitian sebagai berikut:

1. Pengumpulan Data:

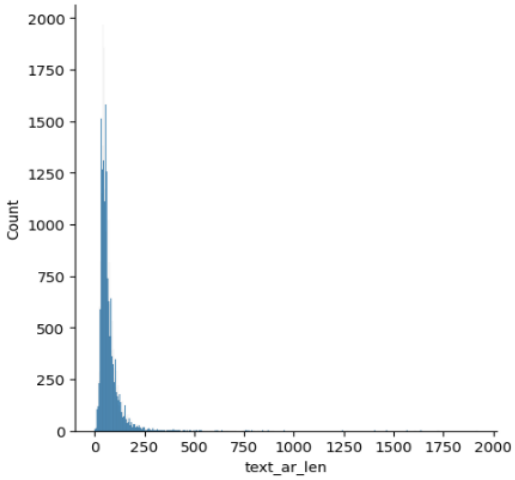
Dataset hadis dapat dilihat pada gambar 2 yang menunjukkan bahwa terdapat kolom id hadis dan kolom isi dari hadis tersebut. Data Hadis bukhari muslim ini terdiri dari 34.441 hadis yang di masing masing kolom isi terdiri dari sanad dan matan di setiap hadisnya. Dengan panjang hadis bervariasi sesuai dengan jumlah kata pada sanat maupun jumlah kata pada sanadnya.

	id	hadith_id	text_ar
0	0	1	...حدثنا الحميدي عبد الله بن الزبير، قال حدثنا سف
1	1	2	...حدثنا عبد الله بن يوسف، قال أخبرنا مالك، عن هش
2	2	3	... «حدثنا يحيى بن بكير، قال حدثنا الليث، عن عقيل
3	3	4	...قال ابن شهاب وأخبرني أبو سلمة بن عبد الرحمن، أ
4	4	5	...حدثنا موسى بن إسماعيل، قال حدثنا أبو عوانة، قا
5	5	6	...حدثنا عبدان، قال أخبرنا عبد الله، قال أخبرنا ي
6	6	7	...حدثنا أبو اليمان الحكم بن نافع، قال أخبرنا شعي
7	0	8	...حدثنا عبيد الله بن موسى، قال أخبرنا حنظلة بن أ

Sumber : Hasil Penelitaian (2025)

Gambar 2. Dataset Hadis Bukhori Muslim

Pada gambar 3 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.



Sumber: Hasil Penelitaian (2025)

Gambar 3. Distribusi panjang Teks Hadis

2. Pra-pemrosesan:

Pada gambar 4 menunjukkan hasil dari pra-pemrosesan dengan ada penmbahan kolom panjang kata hadis dengan nama text_ar_len. Pada pra-pemrosesan dataset 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.

	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، أ	87
4	4	5	...حدثنا موسى بن إسماعيل، قال حدثنا أبو عوانة، قا	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 عبيد 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 القرشي، قال حدثنا أ	46
11	4	12	... ،حدثنا عمرو بن خالد، قال حدثنا الليث، عن يزيد	44
12	5	13	...،حدثنا مسدد، قال حدثنا يحيى، عن شعبة، عن قتادة	45
13	6	14	...حدثنا أبو اليمان، قال أخبرنا شعيب، قال حدثنا أ	41
14	7	15	...حدثنا يعقوب بن إبراهيم، قال حدثنا ابن علية، عن	52
15	8	16	...حدثنا محمد بن المثنى، قال حدثنا عبد الوهاب الث	60
16	9	17	...حدثنا أبو الوليد، قال حدثنا شعبة، قال أخبرني ع	35
17	10	18	... ،حدثنا أبو اليمان، قال أخبرنا شعيب، عن الزهري	109
18	11	19	...حدثنا عبد الله بن مسلمة، عن مالك، عن عبد الرحم	53
19	12	20	... ،حدثنا محمد بن سلام، قال أخبرنا عبدة، عن هشام	61

Sumber : Hasil Penelitaian (2025)

Gambar 4. Menambahkan Kolom Panjang Teks

3. Penerapan Model BERT:

Dalam tahapan ini proses yang dilakukan adalah membuat sebuah nilai embeding berupa vektor numerik yang dapat digambarkan pada gambar 5 yang menunjukkan bahwa semakin tinggi nilainya maka semakin penting kata tersebut sebagai representasi topik

```
[('واللفظ', np.float64(0.006159098085086656)),
 ('يعني', np.float64(0.005638201526386275)),
 ('ثم', np.float64(0.005295135260450917)),
 ('فقال', np.float64(0.005128943015187297)),
 ('رسول', np.float64(0.004925790679522881)),
 ('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))]
```

Sumber : Hasil Penelitaian (2025)

Gambar 5. Nilai Vektor Numerik

Hasil dari embeding vektor adalah sejumlah 768 vektor yang memberikan gambaran bahwa setiap hadis direpresntasikan kedalam tiap nilai sepanjang 768 nilai.

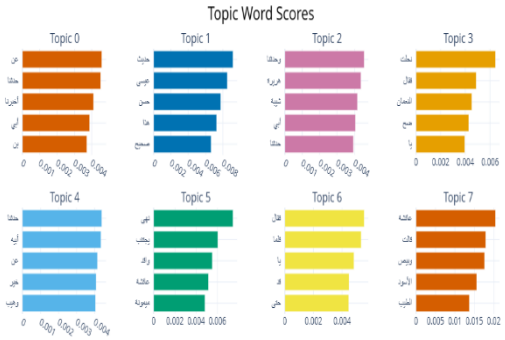
4. Penambahan Fitur Semantik untuk Klusterisasi:

Dalam tahapan ini penambahan fitur semantik berupa panjang teks dan 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.

Setelah menggbungan maka dihasilkan vektor embeding sebesar (34441, 3769) yang mana data tersebut terlalu panjang nilai yang dihasilkan yaitu sebesar 3769 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 100 komponen utama. yang berhasil menampilkan visualisasi dalam bentuk bagan seperti pada gambar 6 yang menyatakan bahwa dapat mengelompokkan hadis dalam 6 topik utama word scores yang mencakup tema-tema sentral seperti sosial, etika, pendidikan, dan ibadah.

5. Analisis dan Interpretasi Hasil:

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

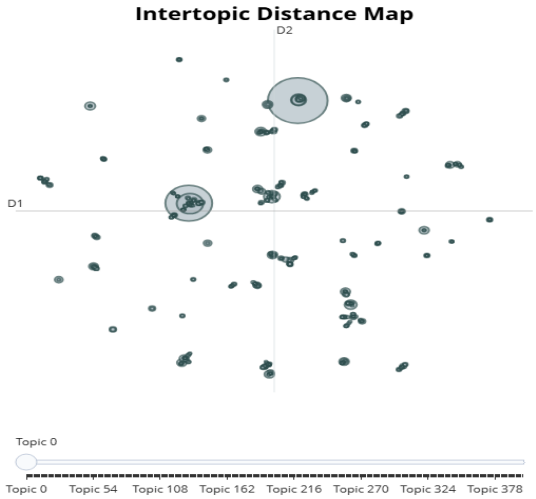


Sumber : Hasil Penelitaian (2025)

Gambar 6. Topik Word Scores

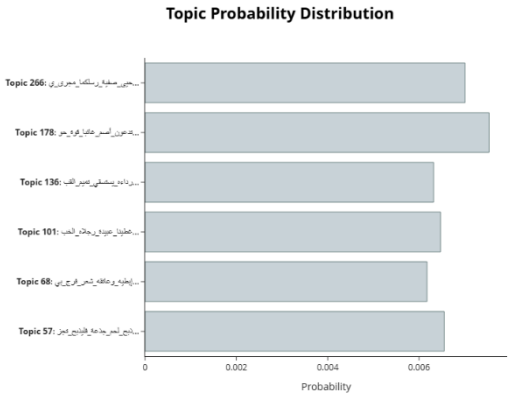
Pada gambar 6 diatas menggambarkan bahwa setiap kluster terdiri atas hadis yang memiliki kesamaan konteks dan makna, yang dihasilkan dari analisis semantik yang lebih mendalam. Hasil ini menarik mengingat BERT mampu menangkap kompleksitas bahasa yang digunakan dalam teks hadis, sehingga kelompok kluster memiliki koherensi yang tinggi berdasarkan studi yang relevan.

Sedangkan pada gambar 7 menggambarkan banyak topik tersebar merata dan ada topik yang saling bertumpuk yang menunjukkan bahwa topik mirip atau redundan. Dan Topik kiri dan kanan atas menunjukkan topik dominan yang menggambarkan bahwa jumlahnya banyak.



Sumber : Hasil Penelitaian (2025)

Gambar 7. Intropic Distance Map

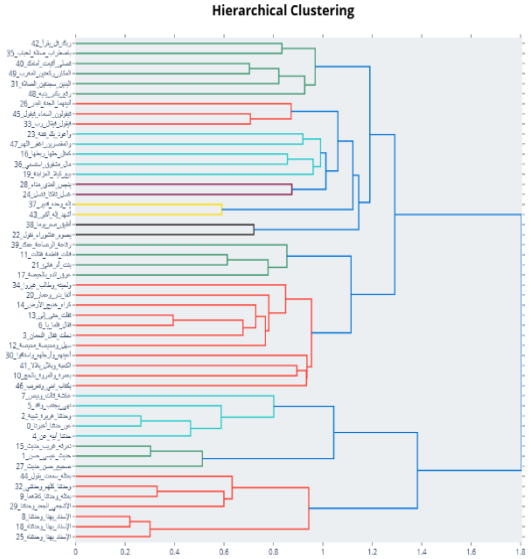


Sumber : Hasil Penelitaian (2025)

Gambar 8. Topik Probability Distribution

Gambar 8 di atas 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.
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 9. Hierarchical Clustering menunjukkan hasil pengelompokan data teks menggunakan metode hierarki klaster (*Hierarchical Clustering*).



Sumber : Hasil Penelitaian (2025)

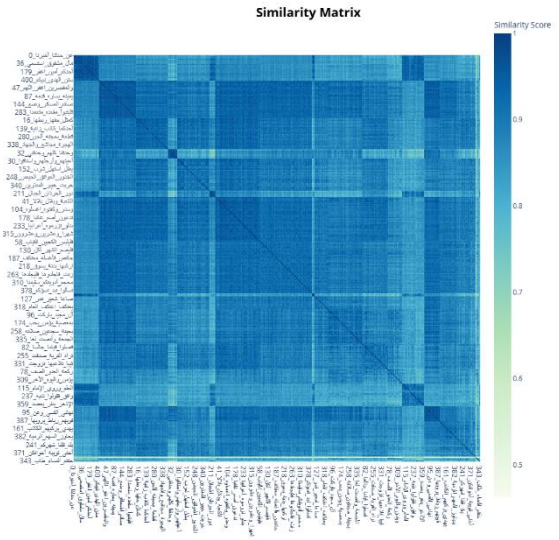
Gambar 9. Hierarichal Clustering

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.
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.
Pada Gambar 10. 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.



Sumber : Hasil Penelitaian (2025)

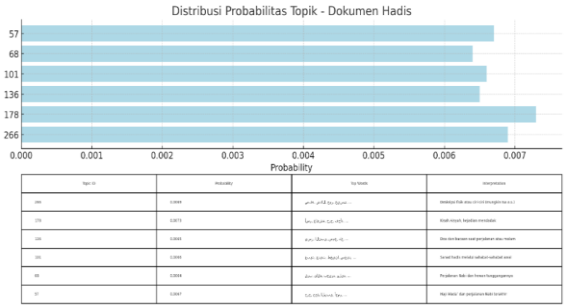
Gambar 10. Similirity Matrix

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).

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.

6. Validasi Model:

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.



Sumber : Hasil Penelitaian (2025)

Gambar 10. Distribusi Probabilitas Topik-Dokumen Hadis

Interpretasi validasi hasil klusterisasi dapat dilihat pada gambar 10 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.

IV. KESIMPULAN

Penelitaian pengembangan model klusterisasi topik hadis Bukhari Muslim menggunakan BERT dengan penambahan fitur semantik menunjukkan bahwa pendekatan ini berhasil meningkatkan akurasi dan interpretabilitas dalam mengelompokkan hadis berdasarkan tema dan konteksnya. Melalui enam tahapan metodologi yang telah dilaksanakan, penelitian ini memberikan gambaran yang lebih jelas mengenai distribusi topik dan hubungan antar hadis yang sebelumnya tidak terlihat.

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.

V. REFERENSI

Aluri, L., & Latha, D. (2023). *HSFO: Hunter Sail Fish Optimizer Enabled Deep Learning for Single Document Abstractive Summarization Based on Semantic Role Labelling for Telugu Text*. https://doi.org/10.21203/rs.3.rs-2889668/v1

Aminah, N., Maryati, M., Bachtiar, M., & Ashpandi. (2025). Hadis Tentang Konsep Manajemen Pengorganisasian Pendidikan Islam Dalam Perspektif Hadits. *Ijis*, *1*(1), 98–106. https://doi.org/10.62567/ijis.v1i1.633

Asy'ari, A. H., Muzakki, M. H., & Hanafi, M. (n.d.). *Clusterization Model of Hadith Topic in Bukhari Muslim Hadith using BERT Algorithm*.

Dodda, R., & Alladi, S. B. (2024). *BERT-based Document Clustering: Unveiling Semantic Patterns in 20News Group, Reuters, and BBC Sports Corpora*. https://doi.org/10.22541/au.171506422.20645846/v1

George, L., & Sumathy, P. (2023). An Integrated Clustering and BERT Framework for Improved Topic Modeling. *International Journal of Information Technology*, *15*(4), 2187–2195. https://doi.org/10.1007/s41870-023-01268-w

Gerritse, E. J. (2022). *Entity-Aware Transformers for Entity Search*. https://doi.org/10.48550/arxiv.2205.00820

Hua, L. (2024). *Integrating Clustering and Semantic Similarity for MAUDE Database Dimensionality Reduction*. https://doi.org/10.1101/2024.12.03.24318439

Imana, Y., Kosasih, E., & Mardi, I. (2024). Madrasah Hadits Dan Sejarah Perkembangannya: Menghubungkan Tradisi Dengan Inovasi Dalam Studi Islam Kontemporer. *Cakrawala*, *1*(2), 141–149. https://doi.org/10.63142/cakrawala.v1i2.68

Li, W. J., Liu, Y., Deng, K., & Wu, X. (2024). *POS-HC: A Part-of-Speech Hierarchical Clustering Approach for Normative Texts Partition*. https://doi.org/10.20944/preprints202402.1575.v1

Liu, T., Yu, H., & Blair, R. H. (2022). Stability Estimation for Unsupervised Clustering: A Review. *Wiley Interdisciplinary Reviews Computational Statistics*, *14*(6). https://doi.org/10.1002/wics.1575

Maulida, F. (2023). The Concept of Political Ethics in Islam (Perspective of Hadith From Sahih Bukhari and Sahih Muslim). *Aqwal Journal of Qur an and Hadis Studies*, *4*(2), 198–212. https://doi.org/10.28918/aqwal.v4i2.1901

Mudding, A. A. (2024). Mengungkap Opini Publik: Pendekatan BERT-based-caused Untuk Analisis Sentimen Pada Komentar Film. *Journal of System and Computer Engineering (Jsce)*, *5*(1), 36–43. https://doi.org/10.61628/jsce.v5i1.1060

Murfi, H., Agung, Y. J., Nurrohmah, S., Satria, Y., Za'in, C., & Rahayu, D. (2024). *Eigenspace-Based Fuzzy C-Means With Large Language Model BERT for Topic Detection*. https://doi.org/10.21203/rs.3.rs-3637575/v1

Riantika, P. A. (2023). Analisis Keutamaan Sedekah Dan Infak Berdasarkan Hadis Yang Diriwayatkan Oleh Imam Bukhari Dan Imam Muslim. *Jurnal Hibrul Ulama Jurnal Ilmu Pendidikan Dan Keislaman*, *5*(2), 76–82. https://doi.org/10.47662/hibrululama.v5i2.522

Rinjani, C. (2021). Metode Reward Dan Punishment Dalam Pendidikan Islam Perspektif Hadis Bukhari Dan Muslim. *Ruhama Islamic Education Journal*, *4*(2), 185–204. https://doi.org/10.31869/ruhama.v4i2.2918

Rohman, F. (2021). Tujuan Pendidikan Islam Pada Hadis-Hadis Populer Dalam Shahihain. *Ta Dibuna Jurnal Pendidikan Islam*, *10*(3), 367. https://doi.org/10.32832/tadibuna.v10i3.5107

Shen, X., Sun, Y., Zhang, C., Yang, C., Qin, Y., Zhang, W., Nan, J., Che, M., & Gao, D. (2024). Double-Target Self-Supervised Clustering With Multi-Feature Fusion for Medical Question Texts. *Peerj Computer Science*, *10*, e2075. https://doi.org/10.7717/peerj-cs.2075

Subakti, A., Murfi, H., & Hariadi, N. (2022). The Performance of BERT as Data Representation of Text Clustering. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00564-9

Yang, Y. (2024). Comparative Analysis of Strategies of Knowledge Distillation on BERT for Text Matching. *Applied and Computational Engineering*, *51*(1), 112–118. https://doi.org/10.54254/2755-2721/51/20241188

Yulianingsih, Y., & Nursihah, A. (2021). Prophetic Parenting: Ide, Spirit Dan Kontekstualisasi Hadis-Hadis Pendidikan

Anak. *(Japra) Jurnal Pendidikan Raudhatul Athfal (Japra)*, *4*(2), 1–18. https://doi.org/10.15575/japra.v4i2.15724

Zhou, Y., Song, C., Li, J., Wu, Z., Bian, Y., Su, D., & Meng, H. (2021). *Enhancing Word-Level Semantic Representation via Dependency Structure for Expressive Text-to-Speech Synthesis*. https://doi.org/10.48550/arxiv.2104.06835