

# Perbandingan Kinerja Algoritma *Vector Space Model* dan *Latent Semantic Indexing* dalam *Retrival* Informasi Pencarian Buku

## *Comparison of the Performance of Vector Space Model and Latent Semantic Indexing Algorithms in Book Search Information Retrieval*

<sup>1</sup>Ida Bagus Satriya Wibawa\*, <sup>2</sup>I Gede Aris Gunadi, <sup>3</sup>I Made Gede Sunarya

<sup>1,2,3</sup>Ilmu Komputer, Pascasarjana, Universitas Pendidikan Ganesha

<sup>1,2,3</sup>Jl. Udayana, Buleleng Bali

\*e-mail: [bagus.satriya@student.undiksha.ac.id](mailto:bagus.satriya@student.undiksha.ac.id), [igedearisgunadi@undiksha.ac.id](mailto:igedearisgunadi@undiksha.ac.id),  
[sunarya@undiksha.ac.id](mailto:sunarya@undiksha.ac.id)

(received: 25 May 2026, revised: 9 June 2026, accepted: 10 June 2026)

### Abstrak

Pemilihan metode yang tepat dalam sistem temu kembali informasi (information retrieval) merupakan faktor krusial untuk menghasilkan pencarian yang akurat dan efisien. Penelitian ini bertujuan membandingkan kinerja algoritma *Vector Space Model* (VSM) dan *Latent Semantic Indexing* (LSI) dalam konteks retrieval informasi pencarian buku, khususnya ditinjau dari kemampuan semantik dan efisiensi komputasi. Dataset yang digunakan merupakan gabungan dua dataset buku dari platform Kaggle yang bersumber dari katalog Amazon, dengan total 133.491 data setelah normalisasi dan penghapusan duplikat. Atribut yang dianalisis mencakup judul dan deskripsi buku. Pengujian dilakukan melalui dua skenario, yaitu pengujian polysemy dan synonymity yang dievaluasi menggunakan Mean Absolute Percentage Error (MAPE) terhadap 20 dokumen dan 5 query, serta pengujian kecepatan retrieval menggunakan time response pada Google Colab dengan 20 variasi volume dataset. Hasil pengujian polysemy dan synonymity menunjukkan bahwa LSI unggul pada tiga dari lima query dengan skor MAPE terbaik sebesar 33,20%, sementara VSM mencatat skor MAPE terendah sebesar 36,35% namun mengalami degradasi hingga 72,01% pada query dengan ambiguitas tinggi. Sebaliknya, pada pengujian kecepatan, VSM menunjukkan efisiensi superior dengan waktu respons berkisar antara 0,4583 ms hingga 7,2597 ms, jauh lebih cepat dibandingkan LSI yang memerlukan 2,5290 ms hingga 27,6777 ms. Kedua algoritma menunjukkan tren peningkatan waktu yang linear dengan koefisien determinasi masing-masing  $R^2=0,999$  dan  $R^2=0,997$ . Penelitian ini menyimpulkan adanya trade-off signifikan antara akurasi semantik dan efisiensi komputasi, di mana LSI lebih unggul dalam pemahaman makna, sedangkan VSM lebih superior dalam kecepatan pemrosesan.

**Kata kunci:** *information retrieval, latent semantic indexing, mean absolute percentage error, polysemy, synonymity, vector space model.*

### Abstract

Selecting an appropriate method for an information retrieval system is a critical factor in achieving accurate and efficient search performance. This study aims to compare the performance of the *Vector Space Model* (VSM) and *Latent Semantic Indexing* (LSI) in the context of book information retrieval, with particular emphasis on semantic capability and computational efficiency. The dataset was constructed by merging two book datasets obtained from the Kaggle platform, which were originally sourced from Amazon's book catalog. After data normalization and duplicate removal, the final dataset consisted of 133,491 book records. The analysis focused on two primary attributes: book titles and book descriptions. The evaluation was conducted through two experimental scenarios: polysemy and synonymity testing, assessed using the Mean Absolute Percentage Error (MAPE) across 20 documents and five search queries, and retrieval speed testing, measured by response time on Google Colab using 20 dataset size variations. The experimental results indicate that LSI outperformed VSM in three of the five search queries, achieving the best MAPE score of 33.20%, whereas VSM recorded its lowest MAPE of 36.35% but deteriorated to 72.01% for queries with high semantic ambiguity. In

<http://sistemasi.ftik.unisi.ac.id>

*contrast, VSM demonstrated superior computational efficiency in the retrieval speed evaluation, with response times ranging from 0.4583 ms to 7.2597 ms, substantially faster than LSI, which required between 2.5290 ms and 27.6777 ms. Both algorithms exhibited a linear increase in response time as the dataset size increased, with coefficients of determination of  $R^2 = 0.999$  for VSM and  $R^2 = 0.997$  for LSI. The findings reveal a significant trade-off between semantic accuracy and computational efficiency: LSI provides superior semantic understanding for information retrieval, whereas VSM offers substantially faster retrieval performance.*

**Keywords:** *information retrieval, latent semantic indexing, mean absolute percentage error, polysemy, synonymity, vector space model.*

## 1 Pendahuluan

Perkembangan teknologi informasi dalam satu dekade terakhir telah memberikan dampak yang sangat signifikan terhadap cara masyarakat memperoleh dan mengakses informasi [1]. Aktivitas pencarian informasi kini menjadi bagian penting dalam kehidupan sehari-hari, baik untuk memperoleh berita terkini, mencari referensi akademik, mempelajari keterampilan baru, hingga menemukan produk atau layanan yang dibutuhkan [2]. Kemudahan akses internet serta meningkatnya jumlah data digital menyebabkan kebutuhan terhadap sistem pencarian informasi yang cepat, akurat, dan relevan menjadi semakin tinggi [3]. Dalam praktiknya, masyarakat umumnya memanfaatkan mesin pencari seperti Google, Bing, dan Yahoo untuk melakukan pencarian berbasis teks secara efisien [4]. Selain itu, fitur pencarian juga telah banyak diterapkan pada platform marketplace, perpustakaan digital, media sosial, maupun sistem e-commerce untuk membantu pengguna menemukan informasi sesuai kebutuhan [5].

Pada proses pencarian informasi, pengguna memasukkan query atau kata kunci tertentu yang kemudian dicocokkan dengan kumpulan dokumen yang tersimpan di dalam sistem [6]. Hasil pencarian yang ditampilkan dapat memiliki tingkat relevansi yang berbeda-beda tergantung pada metode retrieval informasi yang digunakan. Retrieval informasi merupakan cabang ilmu komputer yang berfokus pada proses pencarian, pengumpulan, pengorganisasian, dan penyajian informasi dari sekumpulan dokumen berdasarkan isi maupun konteks dokumen tersebut. Informasi yang dicari direpresentasikan dalam bentuk query yang terdiri atas satu atau lebih term sebagai dasar proses pencairan [7]. Dokumen yang diproses pada retrieval informasi umumnya berupa data semi-terstruktur maupun tidak terstruktur seperti buku, artikel, jurnal ilmiah, berita, dan berbagai file digital lainnya. Dalam sistem retrieval informasi berbasis teks, kemampuan sistem dalam memahami hubungan antar kata sangat menentukan kualitas hasil pencarian. Sebuah query dapat memiliki makna ganda (polysemy) maupun memiliki sinonim yang berbeda tetapi merujuk pada makna yang sama (synonymity). Permasalahan tersebut sering menyebabkan hasil pencarian menjadi kurang relevan apabila metode retrieval yang digunakan tidak mampu memahami hubungan semantik antar kata. Oleh karena itu, pemilihan metode retrieval informasi menjadi faktor penting dalam menghasilkan sistem pencarian yang efektif dan efisien [8].

Salah satu metode retrieval informasi yang paling umum digunakan adalah Vector Space Model (VSM). Metode ini menggunakan model aljabar untuk merepresentasikan dokumen dan query ke dalam bentuk vektor berdasarkan bobot kata menggunakan pendekatan Term Frequency-Inverse Document Frequency (TF-IDF). Selanjutnya, tingkat kemiripan antara query dan dokumen dihitung menggunakan cosine similarity untuk menentukan ranking dokumen yang paling relevan. Penelitian yang dilakukan oleh Angdresey dkk. (2021) menunjukkan bahwa sistem retrieval berbasis VSM mampu menampilkan dokumen paling relevan dengan tingkat similaritas mencapai 78,65%. Penelitian lain oleh Salmon dkk. (2020) juga menunjukkan bahwa metode VSM pada search engine perpustakaan menghasilkan nilai recall dan precision sebesar 1, yang mengindikasikan bahwa sistem mampu menemukan dokumen relevan secara optimal. Meskipun memiliki performa yang baik dalam kecepatan pencarian dan implementasi yang relatif sederhana, metode VSM memiliki keterbatasan dalam memahami hubungan semantik antar kata. VSM bekerja berdasarkan kecocokan term secara langsung sehingga kesulitan dalam menangani kasus polysemy dan synonymity [9]. Menurut Agosti dkk. (2020), VSM tradisional belum mampu mengakomodasi fitur linguistik seperti polisemi yang menyebabkan hasil retrieval menjadi kurang akurat. Penelitian oleh Tang dkk. (2020) juga

menegaskan bahwa VSM mengalami kendala dalam membedakan makna kata tanpa adanya konteks tambahan [10].

Sebagai alternatif, Latent Semantic Indexing (LSI) merupakan metode retrieval informasi yang lebih menekankan pada hubungan semantik antar kata dan dokumen. LSI memanfaatkan proses Singular Value Decomposition (SVD) untuk mereduksi dimensi data dan menemukan makna laten yang tersembunyi di dalam dokumen. Dengan pendekatan tersebut, LSI mampu memahami keterkaitan makna antar kata meskipun kata yang digunakan berbeda secara eksplisit. Penelitian oleh Hasibuan dkk. (2019) menunjukkan bahwa metode LSI mampu menghasilkan tingkat akurasi hingga 79,54% [11]. Penelitian lain oleh Ardiansyah dkk. (2021) juga menunjukkan bahwa LSI menghasilkan recall sebesar 100%, precision sebesar 70%, dan f-measure sebesar 82% pada sistem retrieval informasi. Hasil tersebut menunjukkan bahwa LSI lebih efektif dalam menangani sinonim dan ambiguitas makna dibandingkan VSM. Namun demikian, metode LSI juga memiliki kelemahan, khususnya pada aspek kompleksitas komputasi. Proses reduksi dimensi menggunakan Singular Value Decomposition membutuhkan daya komputasi yang lebih besar dibandingkan VSM, terutama ketika diterapkan pada dataset berukuran besar. Yee (2020) menyatakan bahwa proses komputasi pada LSI relatif lebih lambat karena tingginya kebutuhan pemrosesan matriks [12]. Hal serupa juga disampaikan oleh Mjal (2020), yang menyebutkan bahwa perhitungan jarak antara query dan dokumen pada metode LSI membutuhkan waktu komputasi yang lebih tinggi dibandingkan metode retrieval tradisional [13].

Berdasarkan uraian tersebut, dapat diketahui bahwa Vector Space Model (VSM) dan Latent Semantic Indexing (LSI) memiliki karakteristik, keunggulan, dan kelemahan yang berbeda. VSM unggul dalam kecepatan komputasi dan efisiensi proses retrieval, sedangkan LSI lebih unggul dalam memahami hubungan semantik dan menangani ambiguitas bahasa. Oleh karena itu, diperlukan penelitian komparatif untuk menganalisis performa dari masing – masing metode, baik metode Vector Space Model (VSM) ataupun Latent Semantic Indexing (LSI), serta membandingkan kinerja antara kedua metode tersebut dalam retrieval informasi pencarian buku. Perbandingan tersebut difokuskan pada kemampuan metode dalam menangani polysemy dan synonymity, serta efisiensi waktu retrieval yang dihasilkan [14]. Pada penelitian ini, perbandingan kinerja antara metode VSM dan LSI dilakukan menggunakan dataset buku yang bersumber dari katalog Amazon pada platform Kaggle. Pengujian dilakukan melalui dua skenario utama. Skenario pertama adalah pengujian polysemy dan synonymity untuk mengukur tingkat akurasi metode dalam memahami hubungan semantik menggunakan Mean Absolute Percentage Error (MAPE). Skenario kedua adalah pengujian kecepatan retrieval menggunakan time response pada Google Colab untuk mengevaluasi efisiensi komputasi masing-masing metode. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam menentukan metode retrieval informasi yang paling optimal berdasarkan kebutuhan sistem, baik dari aspek akurasi semantik maupun efisiensi komputasi [15].

## **2 Tinjauan Literatur**

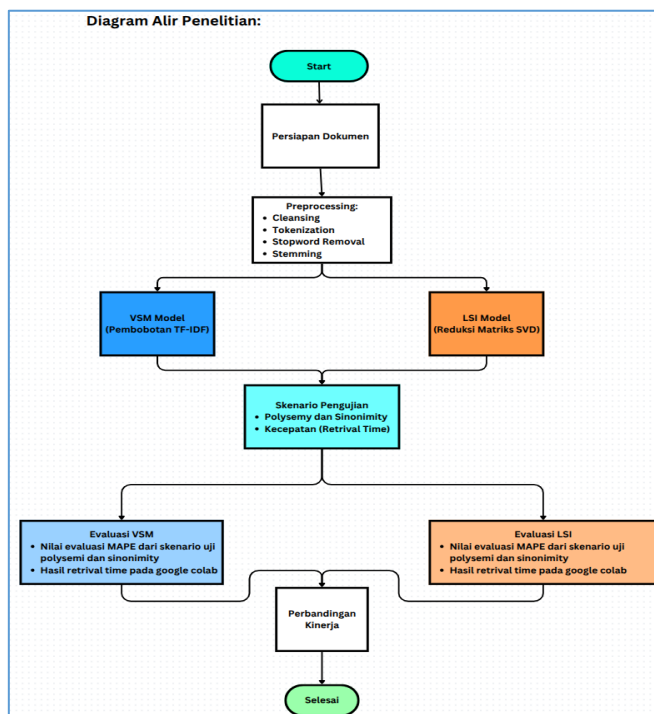
Penelitian mengenai information retrieval telah banyak dilakukan menggunakan berbagai metode untuk meningkatkan relevansi hasil pencarian dokumen. Salah satu metode yang umum digunakan adalah Vector Space Model (VSM), yaitu metode retrieval berbasis representasi vektor menggunakan pembobotan TF-IDF dan perhitungan cosine similarity. Penelitian oleh Angdresey dkk. (2021) menunjukkan bahwa VSM mampu menghasilkan tingkat similaritas hingga 78,65% pada sistem pencarian ayat Alkitab. Penelitian lain oleh Salmon dkk. (2020) juga menunjukkan bahwa VSM menghasilkan nilai recall dan precision sebesar 1 pada search engine perpustakaan, yang menunjukkan kemampuan retrieval dokumen yang sangat baik. Meskipun demikian, beberapa penelitian menunjukkan bahwa VSM memiliki keterbatasan dalam memahami hubungan semantik antar kata. Agosti dkk. (2020) menyatakan bahwa VSM tradisional kurang efektif dalam menangani polysemy karena hanya bergantung pada kecocokan term secara langsung. Tang dkk. (2020) juga menjelaskan bahwa VSM mengalami kesulitan dalam membedakan makna kata ambigu tanpa konteks tambahan [16].

Sebagai alternatif, Latent Semantic Indexing (LSI) dikembangkan untuk meningkatkan kemampuan retrieval berbasis semantik melalui proses Singular Value Decomposition (SVD). Penelitian oleh Hasibuan dkk. (2019) menunjukkan bahwa LSI mampu menghasilkan tingkat akurasi

hingga 79,54%. Selain itu, penelitian Ardiansyah dkk. (2021) menunjukkan bahwa metode LSI menghasilkan recall sebesar 100%, precision sebesar 70%, dan f-measure sebesar 82% pada sistem retrieval informasi dokumen hukum. Hasil tersebut menunjukkan bahwa LSI lebih efektif dalam menangani synonymity dan polysemy dibandingkan VSM. Namun demikian, metode LSI memiliki kelemahan pada aspek efisiensi komputasi. Yee (2020) menyatakan bahwa proses reduksi dimensi pada LSI membutuhkan daya komputasi yang besar sehingga retrieval menjadi lebih lambat pada dataset berskala besar. Hal serupa juga disampaikan oleh Mjal (2020) yang menunjukkan bahwa kompleksitas perhitungan similarity pada LSI lebih tinggi dibandingkan metode retrieval tradisional. Berdasarkan penelitian-penelitian sebelumnya, dapat diketahui bahwa VSM unggul dalam kecepatan komputasi, sedangkan LSI lebih unggul dalam pemahaman semantik. Namun, penelitian yang membandingkan kedua metode tersebut secara khusus pada retrieval informasi pencarian buku dengan pengujian polysemy, synonymity, dan response time masih relatif terbatas. Oleh karena itu, penelitian ini dilakukan untuk membandingkan performa Vector Space Model (VSM) dan Latent Semantic Indexing (LSI) dalam retrieval informasi pencarian buku berdasarkan aspek akurasi semantik dan efisiensi komputasi [17].

### 3 Metode Penelitian

Diagram pada Gambar 1 menggambarkan alur penelitian yang dilakukan untuk menganalisis perbandingan kinerja algoritma Vector Space Model (VSM) dan Latent Semantic Indexing (LSI) dalam retrieval informasi pencarian buku. Tahapan penelitian disusun secara sistematis, dimulai dari pengumpulan dataset, preprocessing data, implementasi kedua algoritma, hingga evaluasi performa untuk menentukan metode yang paling efektif ditinjau dari aspek akurasi semantik dan efisiensi komputasi.



Gambar 1 Diagram alur penelitian

Penelitian ini diawali dengan proses pengumpulan data dari platform Kaggle dengan menggabungkan dua dataset buku yang bersumber dari katalog Amazon, yaitu dataset buku umum dan dataset buku ilmu komputer. Dari atribut yang tersedia, hanya atribut title dan description yang dipilih untuk dianalisis. Setelah melalui proses normalisasi dan penghapusan data duplikat, diperoleh total sebanyak 133.491 dokumen buku. Data yang telah dikumpulkan kemudian melalui tahapan preprocessing untuk meningkatkan kualitas dan konsistensi teks sebelum dilakukan analisis lebih lanjut. Tahapan preprocessing meliputi cleansing, stemming, stopwords removal, dan tokenization. Proses tersebut bertujuan untuk menghapus karakter yang tidak diperlukan, mengubah kata menjadi

bentuk dasarnya, menghilangkan kata-kata yang tidak relevan, serta mengubah teks menjadi token yang terstruktur agar dapat diproses oleh sistem retrieval informasi.

Setelah preprocessing selesai dilakukan, dokumen dianalisis menggunakan dua metode retrieval informasi, yaitu Vector Space Model (VSM) dan Latent Semantic Indexing (LSI). Pada metode VSM, dokumen dan query direpresentasikan dalam bentuk vektor menggunakan pembobotan TF-IDF, kemudian tingkat kemiripan antara query dan dokumen dihitung menggunakan cosine similarity. Sementara itu, pada metode LSI dibentuk matriks term-document yang kemudian direduksi menggunakan Singular Value Decomposition (SVD) untuk menemukan hubungan semantik laten antara kata dan dokumen. Evaluasi performa kedua metode dilakukan menggunakan dua skenario pengujian. Skenario pertama adalah pengujian polysemy dan synonymity yang bertujuan mengevaluasi kemampuan pemahaman semantik masing-masing metode menggunakan Mean Absolute Percentage Error (MAPE) sebagai metrik evaluasi. MAPE digunakan karena mampu mengukur tingkat kesalahan hasil prediksi ranking terhadap ranking aktual dalam bentuk presentase yang memudahkan interpretasi dan perbandingan performa antara metode VSM dan LSI. Semakin kecil nilai MAPE yang diperoleh, maka semakin tinggi tingkat kedekatan hasil retrieval terhadap ranking referensi yang digunakan dalam pengujian. Pengujian plysemy dan synonymity dilakukan menggunakan 20 dokumen dan 5 query yang dirancang untuk mempresentasikan kondisi ambiguitas dan kesamaan makna pada pencarian buku. Skenario kedua adalah pengujian kecepatan retrieval menggunakan pengukuran response time pada Google Colab untuk mengevaluasi efisiensi komputasi pada berbagai ukuran dataset. Selanjutnya, hasil evaluasi dari kedua skenario pengujian tersebut dibandingkan untuk menentukan metode yang paling efektif dan efisien dalam retrieval informasi pencarian buku, khususnya dalam menangani ambiguitas semantik dan performa komputasi.

## 4 Hasil dan Pembahasan

### Hasil

#### 1. Karakteristik Dataset dan Hasil Preprocessing

Penelitian ini menggunakan dua dataset yang diperoleh dari platform Kaggle, yaitu dataset buku ilmu komputer dan dataset buku umum yang bersumber dari katalog Amazon. Dataset pertama terdiri dari 387 data, sedangkan dataset kedua terdiri dari 143.962 data setelah dilakukan penghapusan nilai null. Kedua dataset tersebut kemudian digabungkan dengan menggunakan atribut *title* dan *description* sebagai fokus utama analisis seperti pada Gambar 2. Setelah melalui proses normalisasi dan penghapusan data duplikat, jumlah akhir dataset yang digunakan dalam penelitian ini adalah sebanyak 133.491 dokumen.

155:451 dokumen...

```
data_cleaned1.head()
```

|   | title                            | description                                       | author                  | isbn10     | isbn13         | publish_date | edition | best_seller | top_rated | rating .. |
|---|----------------------------------|---------------------------------------------------|-------------------------|------------|----------------|--------------|---------|-------------|-----------|-----------|
| 0 | The Staff Engineer's Path        | A Guide for Individual Contributors Navigating... | Tanya Reilly            | 1098118731 | 978-1098118730 | 2022-10-25   | 1       | Yes         | No        | 4.7       |
| 1 | Cracking the Coding Interview    | 189 Programming Questions and Solutions           | Gayle Laakmann McDowell | 984782869  | 978-0984782867 | 2015-07-01   | 6       | Yes         | No        | 4.7       |
| 2 | Python Crash Course, 3rd Edition | A Hands-On, Project-Based Introduction to Prog... | Eric Matthes            | 1719502702 | 978-1719502703 | 2023-01-10   | 3       | Yes         | No        | 4.8       |
| 3 | The Pragmatic Programmer         | 20th Anniversary Edition, 2nd Edition, Your Jo... | David Thomas            | NaN        | 978-0133967069 | 2019-09-13   | 2       | Yes         | No        | 4.8       |
| 4 | Clean Code                       | A Handbook of Agile Software Craftsmanship        | Robert C. Martin        | NaN        | 978-0132350884 | 2008-08-01   | 1       | Yes         | No        | 4.7       |

```
data_cleaned2.head()
```

|   | Title                                             | description                                       | authors                | image                                              | previewLink                                            | publisher                  |
|---|---------------------------------------------------|---------------------------------------------------|------------------------|----------------------------------------------------|--------------------------------------------------------|----------------------------|
| 1 | Dr. Seuss: American Icon                          | Philip Nel takes a fascinating look into the k... | [Philip Nel]           | http://books.google.com/books/content?id=lyvHQ...  | http://books.google.nl/books?id=lyvHQscn_jpC&pgC&pg... | A&C Black                  |
| 2 | Wonderful Worship in Smaller Churches             | This resource includes twelve principles in un... | [David R. Ray]         | http://books.google.com/books/content?id=2tsDA...  | http://books.google.nl/books?id=2tsDAAAACAAJ&id...     | NaN                        |
| 3 | Whispers of the Wicked Saints                     | Julia Thomas finds her life spinning out of co... | [Veronica Haddon]      | http://books.google.com/books/content?id=aRSIlg... | http://books.google.nl/books?id=aRSIlgJigJwC&id...     | iUniver                    |
| 5 | The Church of Christ: A Biblical Ecclesiology ... | In The Church of Christ: A Biblical Ecclesiolo... | [Everett Ferguson]     | http://books.google.com/books/content?id=kVqRa...  | http://books.google.nl/books?id=kVqRaPxd8C&pg...       | Wm. B. Eerdmans Publishing |
| 6 | Saint Hyacinth of Poland                          | The story for children 10 and up of St. Hyacin... | [Mary Fabyan Windeatt] | http://books.google.com/books/content?id=imLqA...  | http://books.google.nl/books?id=imLqAAAACAAJ&id...     | Tan Books & Pu             |

```
combined_data_no_duplicates.head()
```

|   | title                                             | description                                       |
|---|---------------------------------------------------|---------------------------------------------------|
| 0 | Dr. Seuss: American Icon                          | Philip Nel takes a fascinating look into the k... |
| 1 | Wonderful Worship in Smaller Churches             | This resource includes twelve principles in un... |
| 2 | Whispers of the Wicked Saints                     | Julia Thomas finds her life spinning out of co... |
| 3 | The Church of Christ: A Biblical Ecclesiology ... | In The Church of Christ: A Biblical Ecclesiolo... |
| 4 | Saint Hyacinth of Poland                          | The story for children 10 and up of St. Hyacin... |

Gambar 2 Gambaran gabungan dataset

Sebelum proses analisis dilakukan, data terlebih dahulu melalui tahapan preprocessing yang terdiri dari cleansing, stemming, stopwords removal, dan tokenization. Tahapan-tahapan tersebut bertujuan untuk menghasilkan data yang lebih bersih, terstruktur, dan konsisten sehingga dapat meningkatkan kualitas retrieval informasi seperti pada Gambar 3.

| books_data[['clean_title', 'clean_description']].head(5)        |                                                   | books_data[['stemmed_title', 'stemmed_description']].head(5)    |  |
|-----------------------------------------------------------------|---------------------------------------------------|-----------------------------------------------------------------|--|
| clean_title                                                     |                                                   | clean_description                                               |  |
| 0                                                               | dr seuss american icon                            | philip nel takes a fascinating look into the k...               |  |
| 1                                                               | wonderful worship in smaller churches             | this resource includes twelve principles in un...               |  |
| 2                                                               | whispers of the wicked saints                     | julia thomas finds her life spinning out of co...               |  |
| 3                                                               | the church of christ a biblical ecclesiology f... | in the church of christ a biblical ecclesiolog...               |  |
| 4                                                               | saint hyacinth of poland                          | the story for children 10 and up of st hyacint...               |  |
| books_data[['no_stop_title', 'no_stop_description']].head(5)    |                                                   | books_data[['tokens_title', 'tokens_description']].head(5)      |  |
| [nltk_data] Downloading package stopwords to /root/nltk_data... |                                                   | [nltk_data] Downloading package punkt_tab to /root/nltk_data... |  |
| [nltk_data] Package stopwords is already up-to-date!            |                                                   | [nltk_data] Unzipping tokenizers/punkt_tab.zip.                 |  |
| no_stop_title                                                   |                                                   | no_stop_description                                             |  |
| 0                                                               | dr seuss american icon                            | philip nel take fascin look key aspect seuss...                 |  |
| 1                                                               | wonder worship smaller church                     | thi resourc includ twelv principl understand s...               |  |
| 2                                                               | whisper wick saint                                | julia thoma find life spin control death husba...               |  |
| 3                                                               | church christ biblic ecclesiolog today            | church christ biblic ecclesiolog today respect...               |  |
| 4                                                               | saint hyacinth poland                             | stori children 10 st hyacinth dominican plant ...               |  |
| tokens_title                                                    |                                                   | tokens_description                                              |  |
| 0                                                               | [dr, seuss, american, icon]                       | [philip, nel, take, fascin, look, key, aspect, ...]             |  |
| 1                                                               | [wonder, worship, smaller, church]                | [thi, resourc, includ, twelv, principl, unders...               |  |
| 2                                                               | [whisper, wick, saint]                            | [julia, thoma, find, life, spin, control, deat...               |  |
| 3                                                               | [church, christ, biblic, ecclesiolog, today]      | [church, christ, biblic, ecclesiolog, today, r...               |  |
| 4                                                               | [saint, hyacinth, poland]                         | [stori, children, 10, st, hyacinth, dominican, ...]             |  |

Gambar 3 Hasil preprocessing

## 2. Hasil Pengujian Polysemy dan Synonymity

Pengujian polysemy dan synonymity dilakukan menggunakan 20 dokumen dengan atribut *title* dan *description*, serta 5 query pencarian yang mengandung makna ganda (*polysemy*) dan kemiripan kata (*synonymity*). 5 query tersebut dipilih karena memiliki tingkat ambiguitas dan variasi makna yang berbeda, sehingga dapat digunakan untuk melakukan evaluasi kemampuan dari masing – masing algoritma dalam memahami hubungan sematik antar kata. Dokumen yang digunakan berasal dari berbagai bidang seperti ilmu komputer, biologi, sains, dan pendidikan sehingga sistem retrieval dapat diuji secara lebih komprehensif dalam memahami konteks makna kata. 5 query yang dapat dilihat pada Tabel 1. Evaluasi dilakukan menggunakan Mean Absolute Percentage Error (MAPE) dengan membandingkan hasil ranking prediksi algoritma terhadap ranking aktual yang ditentukan oleh ahli linguistik untuk memastikan validitas pengujian.

Tabel 1 Daftar 5 query

|         |                                                                                              |
|---------|----------------------------------------------------------------------------------------------|
| Query 1 | <i>computational models and data-driven learning across education, language, and biology</i> |
| Query 2 | <i>network systems and information flow in computing, biology, and communication</i>         |
| Query 3 | <i>semantic models for language understanding and information retrieval</i>                  |
| Query 4 | <i>web-based learning platforms and digital communication systems</i>                        |
| Query 5 | <i>bioinformatics data analysis for genomic and molecular systems</i>                        |

Pada algoritma Vector Space Model (VSM), hasil pengujian menunjukkan nilai MAPE yang bervariasi pada setiap query. Query 1 menghasilkan nilai MAPE sebesar 38,50%, Query 2 sebesar 72,01%, Query 3 sebesar 36,35%, Query 4 sebesar 39,16%, dan Query 5 sebesar 40,96%. Hasil terbaik diperoleh pada Query 3 karena memiliki nilai MAPE paling rendah, sedangkan performa terburuk terjadi pada Query 2 dengan nilai MAPE tertinggi. Hal ini menunjukkan bahwa VSM bekerja cukup baik dalam menangani query dengan tingkat ambiguitas rendah, namun mengalami

penurunan akurasi pada query dengan tingkat polysemy tinggi karena metode ini hanya mengandalkan kemunculan term tanpa memahami hubungan semantik antar kata.

Sementara itu, pada algoritma Latent Semantic Indexing (LSI), hasil pengujian menunjukkan performa yang relatif lebih baik dalam memahami hubungan semantik antar kata. Nilai MAPE yang diperoleh adalah sebesar 33,20% untuk Query 1, 68,61% untuk Query 2, 45,27% untuk Query 3, 35,05% untuk Query 4, dan 48,50% untuk Query 5. Hasil terbaik diperoleh pada Query 1 dengan nilai MAPE terendah sebesar 33,20%, sedangkan performa terburuk juga terjadi pada Query 2. Secara umum, LSI menunjukkan performa yang lebih baik dibandingkan VSM pada sebagian besar query karena mampu menangkap makna laten melalui proses Singular Value Decomposition (SVD), sehingga lebih efektif dalam menangani synonymity dan polysemy.

Penelitian ini juga menunjukkan bahwa jumlah dimensi pada metode LSI sangat memengaruhi performa retrieval. Berdasarkan pengujian jumlah topik dari 5 hingga 10 dimensi, hasil terbaik diperoleh pada penggunaan 9 dimensi dengan nilai MAPE sebesar 33,20%. Hal ini menunjukkan bahwa pemilihan jumlah dimensi yang tepat sangat penting untuk mencapai keseimbangan antara akurasi semantik dan efisiensi komputasi. Jika jumlah dimensi terlalu kecil, sebagian informasi penting dapat hilang, sedangkan jumlah dimensi yang terlalu besar dapat meningkatkan noise pada proses retrieval.

### 3. Hasil Pengujian Kecepatan

Pengujian kecepatan dilakukan untuk membandingkan efisiensi waktu retrieval antara algoritma Vector Space Model (VSM) dan Latent Semantic Indexing (LSI). Setiap algoritma diuji sebanyak 20 kali menggunakan 20 variasi volume dataset yang berbeda, mulai dari 6.675 hingga 133.491 dokumen seperti pada Tabel 2. Evaluasi dilakukan menggunakan *time response* pada IDE Google Colab untuk mengukur waktu yang dibutuhkan oleh masing-masing algoritma dalam melakukan proses retrieval. Pengujian ini bertujuan untuk mengetahui tingkat kestabilan dan efisiensi performa kedua metode seiring dengan peningkatan jumlah data.

**Tabel 2 Daftar volume dataset untuk pengujian kecepatan**

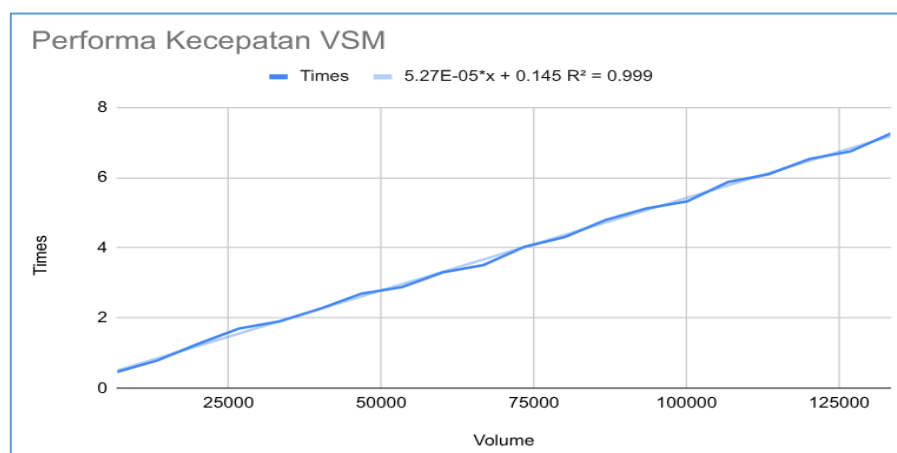
| No. Pengujian | Volume Data (Baris) | No. Pengujian | Volume Data (Baris) |
|---------------|---------------------|---------------|---------------------|
| 1             | 6.675               | 11            | 73.420              |
| 2             | 13.349              | 12            | 80.095              |
| 3             | 20.024              | 13            | 86.769              |
| 4             | 26.698              | 14            | 93.444              |
| 5             | 33.373              | 15            | 100.118             |
| 6             | 40.047              | 16            | 106.793             |
| 7             | 46.722              | 17            | 113.467             |
| 8             | 53.396              | 18            | 120.142             |
| 9             | 60.071              | 19            | 126.816             |
| 10            | 66.746              | 20            | 133.491             |

Pada algoritma Vector Space Model (VSM), hasil pengujian menunjukkan bahwa waktu retrieval meningkat secara proporsional seiring bertambahnya ukuran dataset. Pada Tabel 3 volume data terkecil sebesar 6.675 dokumen, VSM membutuhkan waktu sebesar 0,4583 ms, sedangkan pada volume data terbesar sebesar 133.491 dokumen, waktu retrieval meningkat menjadi 7,2597. Grafik performa pada Gambar 4 menunjukkan tren linear yang sangat kuat dengan nilai koefisien determinasi ( $R^2$ ) sebesar 0,999. Hal ini menunjukkan bahwa VSM memiliki performa retrieval yang sangat stabil, cepat, dan efisien, bahkan ketika digunakan pada dataset berskala besar. Proses tersebut didukung oleh pembobotan TF-IDF dan perhitungan cosine similarity yang secara komputasi relatif sederhana.

**Tabel 3 Hasil pengujian kecepatan vector space model (VSM)**

| Volume Data | Time Response (ms) | Volume Data | Time Response (ms) |
|-------------|--------------------|-------------|--------------------|
| 6.675       | 0,4583             | 73.420      | 4,0246             |
| 13.349      | 0,7870             | 80.095      | 4,3048             |
| 20.024      | 1,2589             | 86.769      | 4,7954             |
| 26.698      | 1,6981             | 93.444      | 5,1248             |

|               |        |         |        |
|---------------|--------|---------|--------|
| <b>33.373</b> | 1,9020 | 100.118 | 5,3228 |
| <b>40.047</b> | 2,2665 | 106.793 | 5,8747 |
| <b>46.722</b> | 2,6928 | 113.467 | 6,0956 |
| <b>53.396</b> | 2,8794 | 120.142 | 6,5362 |
| <b>60.071</b> | 3,3000 | 126.816 | 6,7482 |
| <b>66.746</b> | 3,5060 | 133.491 | 7,2597 |

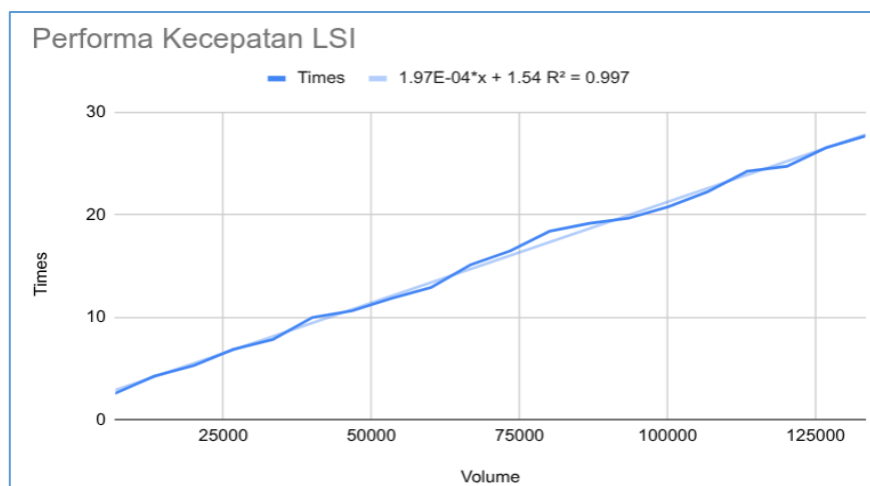


**Gambar 4 Performa kecepatan vector space model (VSM)**

Sementara itu, pada algoritma Latent Semantic Indexing (LSI), hasil pengujian menunjukkan waktu retrieval yang lebih lama dibandingkan VSM. Pada Tabel 4 volume data terkecil LSI membutuhkan waktu sebesar 2,5290 ms, dan meningkat menjadi 27,6777 ms pada volume data terbesar. Grafik performa pada Gambar 5 juga menunjukkan tren linear yang stabil dengan nilai koefisien determinasi ( $R^2$ ) sebesar 0,997. Meskipun LSI masih menunjukkan skalabilitas yang baik, proses retrieval membutuhkan waktu yang lebih lama karena adanya tahapan dekomposisi matriks menggunakan Singular Value Decomposition (SVD) untuk menangkap hubungan semantik laten antara kata dan dokumen. Kompleksitas komputasi tersebut menjadi faktor utama yang menyebabkan LSI lebih lambat dibandingkan VSM.

**Tabel 4 Hasil pengujian kecepatan latent semantic indexing (LSI)**

| <b>Volume Data</b> | <b>Time Response (ms)</b> | <b>Volume Data</b> | <b>Time Response (ms)</b> |
|--------------------|---------------------------|--------------------|---------------------------|
| <b>6.675</b>       | 2,5290                    | 73.420             | 16,4484                   |
| <b>13.349</b>      | 4,2244                    | 80.095             | 18,3776                   |
| <b>20.024</b>      | 5,2786                    | 86.769             | 19,1464                   |
| <b>26.698</b>      | 6,8440                    | 93.444             | 19,6514                   |
| <b>33.373</b>      | 7,8110                    | 100.118            | 20,7565                   |
| <b>40.047</b>      | 9,9728                    | 106.793            | 22,2437                   |
| <b>46.722</b>      | 10,6166                   | 113.467            | 24,2486                   |
| <b>53.396</b>      | 11,8289                   | 120.142            | 24,7159                   |
| <b>60.071</b>      | 12,8864                   | 126.816            | 26,5426                   |
| <b>66.746</b>      | 15,1111                   | 133.491            | 27,6777                   |



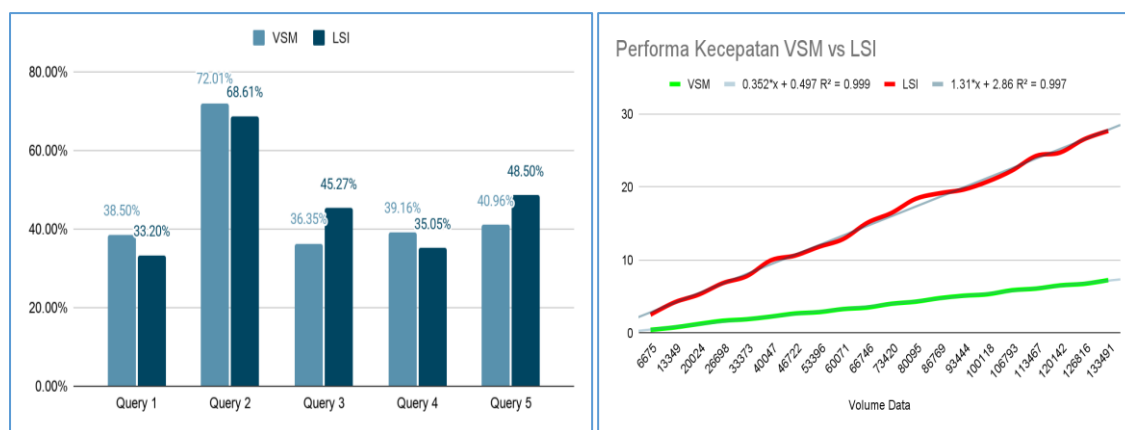
**Gambar 5 Performa kecepatan *latent semantic indexing* (LSI)**

Berdasarkan hasil pengujian tersebut, dapat disimpulkan bahwa VSM lebih unggul dalam aspek kecepatan retrieval, sedangkan LSI membutuhkan waktu komputasi yang lebih besar karena proses analisis semantik yang lebih kompleks. Hal ini menunjukkan adanya trade-off antara kecepatan dan pemahaman semantik, di mana VSM lebih cocok digunakan pada sistem yang mengutamakan efisiensi waktu, sementara LSI lebih sesuai digunakan ketika akurasi semantik menjadi prioritas utama.

### Pembahasan

Berdasarkan hasil perbandingan pada Gambar 6, Vector Space Model (VSM) dan Latent Semantic Indexing (LSI) menunjukkan keunggulan yang berbeda dalam performa retrieval. Pada skenario pengujian polysemy dan synonymity, VSM menghasilkan performa yang bervariasi pada lima query, dengan hasil terbaik pada Query 3 sebesar 36,35% MAPE dan hasil terburuk pada Query 2 sebesar 72,01% MAPE. Hal ini menunjukkan bahwa VSM bekerja cukup baik pada query dengan tingkat ambiguitas rendah, namun memiliki keterbatasan dalam menangani polysemy dan synonymity karena hanya mengandalkan pembobotan TF-IDF dan cosine similarity tanpa memahami hubungan semantik antar kata. Sebaliknya, LSI menunjukkan performa yang lebih baik pada sebagian besar query dan berhasil mengungguli VSM pada tiga dari lima query yang diuji. Hasil terbaik diperoleh pada Query 1 dengan nilai MAPE sebesar 33,20%, sedangkan tingkat kesalahan tertinggi terjadi pada Query 2 sebesar 68,61%. Tingginya nilai MAPE pada Query 2 menunjukkan bahwa VSM mengalami kesulitan dalam menangani istilah yang memiliki makna ganda pada berbagai konteks. Hasil ini mendukung penelitian Agosti dkk. (2020) yang menyatakan bahwa pendekatan *Vector Space Model* tradisional masih memiliki keterbatasan dalam mengakomodasi fenomena polysemy karena proses retrieval sangat bergantung pada kemunculan term secara eksplisit. Performa LSI yang lebih baik didukung oleh proses Singular Value Decomposition (SVD) yang mampu menangkap hubungan semantik laten serta mengurangi noise pada proses retrieval. Dengan penggunaan 9 dimensi, LSI menghasilkan performa paling optimal dan menunjukkan tingkat akurasi yang lebih baik dalam menangani ambiguitas semantik. Hasil ini sejalan dengan penelitian Hasibuan dkk. (2019) dan Ardiansyah dkk. (2021) yang menyatakan bahwa LSI lebih efektif dalam menangani hubungan semantik antar kata karena mampu mempresentasikan makna laten dokumen melalui proses Singular Value Decomposition (SVD). Kemampuan tersebut menyebabkan LSI lebih adaptif dalam menghadapi kasus polysemy dan sunonymity dibandingkan metode berbasis pencocokan term secara langsung seperti VSM. Untuk memperkuat hasil perbandingan, dilakukan uji statistik menggunakan metode Wilcoxon Signed-Rank Test terhadap nilai MAPE yang diperoleh terhadap 5 query pengujian. Uji ini dipilih karena jumlah data yang kecil dan berupa pasangan hasil pengukuran dari metode VSM dan LSI. Hasil pengujian menghasilkan nilai sebesar 0,183 ( $p > 0,5$ ), sehingga tidak terdapat perbedaan yang signifikan secara statistik antara performa VSM dan LSI berdasarkan nilai MAPE. Meskipun demikian, secara deskriptif LSI menunjukkan performa yang lebih baik pada 3 dari 5 query yang diujikan. Maka dapat dikatakan metode LSI tetap menunjukkan kecenderungan lebih unggul dalam menangani hubungan semantik dibandingkan VSM.

Dalam aspek kecepatan retrieval, VSM menunjukkan performa yang lebih cepat dan efisien dibandingkan LSI. Waktu retrieval meningkat dari 0,4583 ms hingga 7,2597 ms seiring bertambahnya ukuran dataset, dengan tren linear yang sangat kuat ( $R^2 = 0,999$ ), yang menunjukkan tingkat stabilitas dan efisiensi yang tinggi pada sistem berskala besar. Sementara itu, LSI membutuhkan waktu pemrosesan yang lebih lama, yaitu berkisar antara 2,5290 ms hingga 27,6777 ms, meskipun tetap menunjukkan skalabilitas yang stabil dengan nilai  $R^2 = 0,997$ . Waktu respons yang lebih lama tersebut disebabkan oleh proses *Singular Value Decomposition* (SVD) yang membutuhkan komputasi lebih kompleks. Temuan ini konsisten dengan temuan penelitian Yee (2020) dan Mjal (2020) yang menyatakan bahwa proses reduksi dimensi menggunakan *Singular Value Decomposition* (SVD) membutuhkan sumber daya komputasi yang lebih besar dibandingkan metode retrieval berbasis pencocokan term. Oleh karena itu, meskipun LSI mampu meningkatkan kualitas pemahaman semantik, metode ini cenderung memiliki waktu pemrosesan yang lebih tinggi dibandingkan VSM. Secara keseluruhan, VSM lebih sesuai digunakan pada sistem yang mengutamakan kecepatan dan efisiensi, sedangkan LSI lebih cocok digunakan pada sistem yang membutuhkan pemahaman semantik yang lebih mendalam serta tingkat akurasi retrieval yang lebih tinggi.



Gambar 6 Perbandingan nilai MAPE dan performa kecepatan antara VSM dan LSI

## 5 Kesimpulan

Penelitian ini bertujuan untuk membandingkan kinerja algoritma Vector Space Model (VSM) dan Latent Semantic Indexing (LSI) dalam retrieval informasi pencarian buku menggunakan 133.491 dokumen dari katalog Amazon. Pengujian dilakukan melalui dua skenario utama, yaitu pengujian polysemy dan synonymity serta pengujian kecepatan retrieval. Hasil pengujian polysemy dan synonymity menunjukkan bahwa Latent Semantic Indexing (LSI) memiliki performa yang lebih baik dibandingkan Vector Space Model (VSM) dalam memprediksi ranking dokumen. Berdasarkan nilai Mean Absolute Percentage Error (MAPE), LSI menghasilkan skor terbaik sebesar 33,20%, sedangkan VSM mengalami tingkat kesalahan tertinggi sebesar 72,01%. Keunggulan LSI didukung oleh proses Singular Value Decomposition (SVD) yang mampu menangkap hubungan semantik laten antar kata, sementara VSM hanya mengandalkan pembobotan TF-IDF dan cosine similarity. Pada pengujian kecepatan, VSM menunjukkan performa yang lebih cepat dan stabil dibandingkan LSI. Waktu retrieval VSM berkisar antara 0,4583 ms hingga 7,2597 ms, sedangkan LSI membutuhkan waktu antara 2,5290 ms hingga 27,6777 ms. Hal ini disebabkan karena proses SVD pada LSI membutuhkan komputasi yang lebih kompleks. Penelitian ini menyimpulkan bahwa terdapat trade-off antara akurasi semantik dan efisiensi komputasi. LSI lebih cocok digunakan untuk sistem yang memprioritaskan akurasi dan pemahaman makna, sedangkan VSM lebih sesuai untuk sistem yang membutuhkan kecepatan retrieval dan efisiensi pada dataset berskala besar.

## Referensi

- [1] K. G. Segara and M. I. P. Nasution, "Perkembangan Teknologi Informasi di Indonesia : Tantangan dan Peluang," *J. Sains Student Res.*, Vol. 3, No. 1, 2025, DOI: <https://doi.org/10.61722/jssr.v3i1.3128>.

- [1] N. Agosti, G. M. Di Nunzio, dan M. Ferro, "A Study of Polysemy in Information Retrieval Systems," *Information Processing and Management*, Vol. 57, No. 6, pp. 1–12, 2020. DOI: 10.1016/j.ipm.2020.102305
- [2] S. Angdresey, E. Pawan, dan F. Liantoni, "Information Retrieval System in The Bible using Vector Space Model," *Journal of Physics: Conference Series*, Vol. 1783, No. 1, pp. 1–9, 2021. DOI: 10.1088/1742-6596/1783/1/012040
- [3] A. H. Ardiansyah, A. R. Atmadja, dan I. M. A. Suyadnya, "Penerapan Latent Semantic Indexing pada Sistem Temu Balik Informasi pada Undang-Undang Pemilu berdasarkan Kasus," *Jurnal Teknologi Informasi dan Ilmu Komputer*, Vol. 8, No. 5, pp. 987–996, 2021. DOI: 10.25126/jtiik.2021855472
- [4] D. Darmawan, N. Syamsyiah, A. A. Alhasna, dan A. Wafi, "Telaah Pustaka Peran Literasi Digital dalam Membangun Daya Pikir Kritis Mahasiswa Masa Kini," *Jurnal Ilmiah Multidisiplin*, Vol. 1, No. 4, pp. 1195–1205, 2025. DOI: 10.63822/k8qdp29
- [5] M. S. Hasibuan, R. Rahim, dan Mesran, "Model Detecting Learning Styles with Artificial Neural Network and Latent Semantic Indexing," *Journal of Physics: Conference Series*, Vol. 1361, No. 1, pp. 1–7, 2019. DOI: 10.1088/1742-6596/1361/1/012018
- [6] T. Indriyatmoko, A. Forbenius, M. Santosa, dan M. Diningrat, "Strategi Optimisasi Kata Kunci dalam meningkatkan Visibilitas di Mesin Pencari," *Information Technology Journal*, Vol. 6, No. 1, pp. 66–72, 2024. DOI: 10.24076/intechnojournal.2024v6i1.1693
- [7] D. Kristanto, R. A. Ramadhani, dan A. B. Setiawan, "Pengoptimalan Sistem Pencarian Repository berbasis Website dengan Cosine Similarity," *Seminar Inovasi dan Teknologi*, Vol. 9, pp. 1331–1340, 2025. [Online]. Available: <https://proceeding.unpkediri.ac.id/index.php/inotek/>
- [8] S. Z. Mjal, "Computational Complexity Analysis of Latent Semantic Indexing in Information Retrieval," *International Journal of Scientific and Technology Research*, Vol. 9, No. 4, pp. 1456–1462, 2020. DOI: 10.13140/RG.2.2.17644.18567
- [9] A. Nabillah dan H. A. Ranggadara, "Mean Absolute Percentage Error for Evaluation of Prediction Models," *Journal of Physics: Conference Series*, Vol. 1641, No. 1, pp. 1–6, 2020. DOI: 10.1088/1742-6596/1641/1/012012
- [10] S. Salmon, M. A. Fauzi, dan R. Kurniawan, "Implementasi Metode Vector Space Model pada Search Engine Perpustakaan," *Jurnal Teknologi Informasi dan Ilmu Komputer*, Vol. 7, No. 6, pp. 1151–1158, 2020. DOI: 10.25126/jtiik.2020762450
- [11] K. G. Segara dan M. I. P. Nasution, "Perkembangan Teknologi Informasi di Indonesia: Tantangan dan Peluang," *Jurnal Sains Student Research*, Vol. 3, No. 1, 2025. DOI: 10.61722/jssr.v3i1.3128
- [12] M. A. Subagyo, R. A. Anya, S. Putri, R. Tanvert, dan T. J. Santosa, "Perancangan Situs Web untuk membandingkan Harga Produk Termurah dari berbagai Platform E-commerce," *Jurnal Teknologi Informasi*, Vol. 11, No. 2, 2025. DOI: 10.52643/jti.v11i2.7366
- [13] S. Sukri, N. A. Samsudin, E. Fadzrin, S. K. A. Khalid, dan L. Trisnawati, "Word2vec-based Latent Semantic Indexing (Word2Vec-LSI) for Contextual Analysis in Job-Matching Application," *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 3, pp. 699–708, 2024. DOI: 10.14569/IJACSA.2024.0150371
- [14] Y. Tang, X. Liu, dan H. Chen, "Improving Vector Space Model for Polysemy Retrieval in Information Search," *IEEE Access*, Vol. 8, pp. 184211–184220, 2020. DOI: 10.1109/ACCESS.2020.3029548
- [15] L. Widaningrum, D. Riana, dan A. Prihandono, "Implementation of TF-IDF Weighting in Text-based Information Retrieval Systems," *Journal of Big Data and Artificial Intelligence*, Vol. 2, No. 1, pp. 45–53, 2022. DOI: 10.31098/jbdai.v2i1.1021
- [16] C. W. Yee, "Performance Analysis of Latent Semantic Indexing for Large Scale Text Retrieval," *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, No. 5, pp. 8123–8130, 2020. DOI: 10.30534/ijatcse/2020/236952020
- [17] R. Yohana, A. Subandi, dan I. Irsyad, "Pengaruh Penggunaan Sistem Informasi Manajemen berbasis Digital terhadap Efisiensi Administrasi Pendidikan," *Academic Social Studies and Global Citizenship Journal*, Vol. 4, No. 1, pp. 21–29, 2024. DOI: 10.47200/AoSSaGCJ.v4i1.2389.