***PROPOSAL SKRIPSI***

***PERBANDINGAN TEKNIK PENCARIAN LEKSIKAL DAN SEMANTIK UNTUK SISTEM PENCARIAN DOKUMEN PUTUSAN PENGADILAN MAHKAMAH AGUNG REPUBLIK INDONESIA***



**Oleh:**

**M. SHOKHIBUL ANWAR**

**200411100130**

**Dosen Pembimbing 1 : Firdaus Solihin, S.Kom., M.Kom.**

**Dosen Pembimbing 2 :**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2023**

# ABSTRAK

Jumlah dokumen putusan pengadilan yang dipublikasikan pada situs web resmi Mahkamah Agung Republik Indonesia terus bertambah dengan rata-rata 900.000 kasus per-tahun. Karenanya, penerapan sistem pencarian yang efektif diperlukan untuk mempermudah dan mempercepat pencarian informasi relevan dari kasus terdahulu. Kendati demikian metode pencarian yang digunakan situs resmi Mahkamah Agung masih menggunakan metode pencarian tradisional berbasis teknik *boolean* yang menuntut pengguna untuk memiliki keahlian teknis yang tinggi agar dapat menggunakan sistem dengan maksimal. Penelitian ini bermaksud untuk memperbandingkan metode pencarian berbasis leksikal (BM25)dan semantik *(Text Embedding)* untuk menguji efektifitas dan efisiensinya pada sistem pencarian dokumen putusan pengadilan. Teknik pencarian leksikal mengandalkan konsep pencocokan ketat *(Strict Matching)* antarakueri dan isi dokumen untuk mencari dokumen yang relevan*,* sedangkan pencarian semantik mencari makna instrinsik untuk merepresentasikan dokumen sehingga pencocokan lebih fleksibel *(Soft Matching)*. Dengan membandingkan kedua teknik ini diharapkan dapat ditemukan teknik yang lebih efisien untuk pencarian dokumen putusan pengadilan sehingga dapat memudahkan pekerjaan pengguna.

Kata kunci: Information Retrieval, Legal Case Retrieval, BM25, BERT

# DAFTAR ISI

[ABSTRAK i](#_Toc160737644)

[DAFTAR ISI ii](#_Toc160737645)

[DAFTAR GAMBAR iii](#_Toc160737646)

[DAFTAR TABEL iv](#_Toc160737647)

[BAB 1 5](#_Toc160737648)

[1.1 Latar Belakang 5](#_Toc160737649)

[1.2 Rumusan Masalah 7](#_Toc160737650)

[1.2.1 Permasalahan 7](#_Toc160737651)

[1.2.2 Metode Usulan 7](#_Toc160737652)

[1.2.3 Pertanyaan penelitian 7](#_Toc160737653)

[1.3 Tujuan Penelitian 7](#_Toc160737654)

[1.3.1 Tujuan 7](#_Toc160737655)

[1.3.2 Manfaat 8](#_Toc160737656)

[1.4 Batasan masalah 8](#_Toc160737657)

[1.5 Sistematika Penulisan 8](#_Toc160737658)

[BAB 2 9](#_Toc160737659)

[2.1 Information Retrieval 9](#_Toc160737660)

[2.2 Legal Case Retrieval 9](#_Toc160737661)

[2.3 Sparse Vector Retriever 9](#_Toc160737662)

[2.4 Dense Vector Retriever 9](#_Toc160737663)

[2.5 BERT retriever. 9](#_Toc160737664)

[BAB 3 10](#_Toc160737665)

[DAFTAR PUSTAKA 10](#_Toc160737666)

# DAFTAR GAMBAR

# DAFTAR TABEL

# 

**PENDAHULUAN**

## Latar Belakang

Negara Kesatuan Republik Indonesia (NKRI) adalah sebuah negara hukum yang secara eksplisit ditegaskan melalui Undang-Undang Dasar Negara Republik Indonesia Pasal 1 ayat (3)[1]. Dalam penegakan hukum, kekuasaan kehakiman khususnya Mahkamah Agung memiliki peranan penting dalam menyelenggarakan peradilan sebagaimana diatur dalam Undang-Undang no. 48 tahun 2009[2]. Melalui serangkaian proses persidangan dan permusyawaratan hakim, suatu kasus akan dinyatakan selesai setelah sahnya putusan hakim berkekuatan hukum tetap dan tidak ada upaya hukum lanjutan[3]. Putusan pengadilan yang sudah berkekuatan hukum tetap memiliki kekuatan mengikat, kekuatan pembuktian dan kekuatan eksekutorial (pelaksanaan) untuk terdakwa atau pihak-pihak yang bersengketa [4].

Putusan pengadilan atau putusan hakim adalah produk dari pemeriksaan perkara di persidangan yang berisikan pernyataan hakim untuk menyelesaikan atau memperlancar pemeriksaan suatu perkara [5]. Putusan hakim terdiri dari banyak macam yang berbeda-beda berdasarkan tiap jenis perkara hukum. Setiap jenis perkara hukum juga memiliki banyak macam putusan yang dibagi berdasarkan beberapa faktor seperti amar putusan, tingkat putusan, tingkat pengakhiran perkara, kehadiran para pihak, dll [5]. Perbedaan-perbedaan jenis ini mempengaruhi karakteristik isi putusan yang mengakibatkan perbedaan fungsi putusan terhadap pihak-pihak yang terlibat. Selain ditujukan bagi pihak-pihak yang terlibat, putusan pengadilan juga dapat berguna untuk dipelajari bagi publik misalnya materi pembelajaran akademisi bidang hukum, atau kalangan aparat dan praktisi hukum.

Data putusan pengadilan yang diunggah pada situs Direktori Putusan MA pada Februari 2024 mencapai lebih dari 8,5 juta dokumen dengan penambahan rata-rata 65 ribu dokumen setiap bulan [6]. Agar bisa mengimbangi banyaknya jumlah dokumen, diperlukan sistem pencarian otomatis yang efektif, efisien dan memiliki skalabilitas baik. Sistem pencarian informasi *(Information Retrieval)* pada bidang hukum khususnya dokumen konten panjang seperti putusan pengadilan memiliki tantangan tersendiri yang berbeda dengan bidang lain. Tantangan ini umumnya berupa pengoptimalan performa terkait besarnya basis data, pemahaman semantik dokumen, pengolahan panjang dokumen, dan pemahaman jargon atau istilah-istilah khusus [7].

Metode pencarian dokumen yang digunakan pada situs resmi Direktori Putusan MA saat penelitian ini dibuat adalah metode pencarian leksikal melalui *Database Indexing* berbasis model pencarian *boolean*. *Boolean search* didasarkan pada logika *Boolean* dan teori himpunan di mana dokumen dan kueri dianggap sebagai *bag-of-words* atau sekumpulan istilah *(terms)* [8]. Cara kerjanya adalah mencari berdasarkan ada atau tidaknya istilah-istilah pada teks kueri pada dokumen, serta terpenuhinya operator logika *boolean* yang ada pada teks kueri. Keunggulan Boolean search ada pada konsepnyayaitu sama persis setiap kata *(hard match)* yang membuatnya mudah diaplikasikan, mudah dipahami dan membuat pengguna memegang kendali penuh atas hasil pencarian [9].

Disisi lain, karakteristik lugas dari metode *boolean retrieval* ini tidak selalu sesuai dengan apa yang dibutuhkan oleh pengguna. Keterbatasan model antara lain adalah pada segi fleksibilitas, sensitivitas terhadap pembuatan kueri, dan tidak adanya pemeringkatan relevansi sehingga menghambat pencarian informasi dengan topik yang luas [10]. Hal ini menuntut pengguna untuk memiliki pengetahuan teknis seperti pemahaman sintaksis dan istilah-istilah pada topik terkait agar dapat menemukan hasil yang tepat secara efektif. Solusi permasalahan-permasalahan tersebut telah banyak diteliti seiring dengan berjalannya waktu dan perkembangan teknologi. Diantaranya dengan implementasi algoritma pemeringkatan *(ranking)*[11], penanganan sinonim dan variasi kata [12], implementasi pemahaman semantik [13], serta pengaplikasian berbagai teknik pemrosesan bahasa alami *(NLP)* [14].

Metode pencarian modern umumnya menggunakan 2 jenis pendekatan yaitu pencarian leksikal dan semantik [15]. Teknik pencarian leksikal didasarkan pada model *Bag of Words* (BOW), di mana kueri dan dokumen direpresentasikan sebagai sekumpulan *term* bersama dengan jumlah kemunculan token yang sama dalam teks contohnya, model *vector space* [16] dan model probabilistik [17]. Sementara itu pencarian semantik merepresentasikan dokumen dengan melakukan teknik *Embedding* yaitu mengubah teksmendadi sebuah vektor padat dengan ukuran tertentu menggunakan teknik *Deep Learning* sehingga makna instrinsik dokumen dapat ditangkap oleh model [18].

Kedua teknik memiliki keunggulan dan kekurangan sesuai dengan karakteristik dan cara kerja masing-masing. Metode leksikal memiliki karakteristik komputasi yang sederhana berdasarkan persamaan kata sehingga memungkinkan waktu pencarian yang lebih cepat dengan akurasi tinggi jika kueri pengguna hampir identik dengan isi dokumen [19]. Dilain sisi, pencarian semantik memiliki keunggulan dalam fleksibilitas karena tidak memerlukan kesamaan persis tiap *term* antara pada kueri dan dokumen, dengan catatan waktu komputasi yang lebih lama [19]. Penelitian ini bermaksud untuk mengukur efektifitas dan efisiensi kedua teknik yang telah disebutkan, dalam sistem pencarian dokumen putusan pengadilan Mahkamah Agung Republik Indonesia. Teknik pencarian semantik berbasis *Pre-trained Transformer* merupakan pendekatan yang lebih baru dan dapat mencapai performa mutakhir pada berbagai bidang pencarian informasi. Namun, dokumen pada bidang hukum seringkali memerlukan kesamaan dan konsistensi istilah-istilah khusus dalam penggunaannya misalnya penomoran atau judul undang-undang, dalam hal ini teknik leksikal seperti BM25 berpotensi lebih unggul daripada teknik pencarian semantik.

## Rumusan Masalah

### Permasalahan

Bagaimana menerapkan sebuah sistem pencarian informasi yang efektif dan efisien untuk mencari dokumen yang relevan dari basis data dokumen putusan pengadilan yang besar dan beragam.

### Metode Usulan

Mencari teknik pencarian terbaik dengan memperbandingkan model pencarian leksikal berbasis *Bag-of-Words:* BM25 (Best Matching 25) dengan teknik pencarian semantik *Dense Vector Embedding* dengan model IndoBERT sebagai *Bi-encoder* untuk mencari metode paling efektif dan efisien untuk pencarian dokumen putusan pengadilan.

### Pertanyaan penelitian

1. Seberapa preci dan efisien algoritma pencarian leksikal berbasis BM25 untuk pencarian dokumen putusan pengadilan?
2. Seberapa efektif dan efisien algoritma pencarian semantik dengan metode Sentence-BERT berbasis model IndoBERT untuk pencarian dokumen putusan pengadilan?

## Tujuan Penelitian

1. **Tujuan**

Memperbandingkan efektifitas dan efisiensi metode pencarian BM25 dengan Sentence-BERT berbasis IndoBERT pada sistem pencarian dokumen putusan pengadilan Mahkamah Agung Republik Indonesia.

1. **Manfaat**

Penelitian ini bertujuan memberikan kontribusi dalam bidang hukum khususnya terkait dokumen putusan pengadilan dengan menciptakan sistem pencarian yang lebih fleksibel dari segi query masukan dan memberikan hasil yang lebih akurat.

## Batasan masalah

Penelitian ini berfokus pada :

1. Data yang digunakan berasal dari <https://github.com/ir-nlp-csui/indo-law> berisi kumpulan dokumen putusan yang sudah dilakukan *parsing*, *cleaning* dan anotasi segmen oleh [20].
2. Jenis putusan yang digunakan hanya jenis putusan perkara pidana umum yang sudah berstatus berkekuatan hukum tetap *(inkracht)*.

## Sistematika Penulisan

Struktur penulisan proposal disusun sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi 5 subbab yaitu: 1) Latar Belakang, 2) Rumusan Masalah, 3) Tujuan Penelitian, 4) Batasan Masalah, dan 5) Sistematika Penulisan. Menjelaskan konteks penelitian seperti pengenalan bidang penelitian, masalah yang ada dan kontribusi penelitian ini pada bidang tersebut.

BAB 2 LANDASAN TEORI

Dalam bab ini dijelaskan tentang teori-teori yang menjadi landasan bagi penelitian ini, yang mana berisikan: 1) *Information Retrieval*, 2.) *Legal* *Case Retrieval*, 3) *Sparse Vector Retriever* 4) *Dense Vector Retriever*, dan 5) *BERT* *retriever*.

BAB 3 METODOLOGI PENELITIAN

Bab ini menjelaskan metode penelitian yang akan diterapkan, mencakup analisis rancangan keseluruhan sistem, Pemrosesan dataset, alur penelitian, dan jadwal penelitian.

# 

**LANDASAN TEORI**

## Information Retrieval

Temu kembali informasi *(Information Retrieval)* adalah salah satu sub-disiplin dari ilmu komputer dengan tujuan menemukan dan mengambil informasi yang relevan sesuai dengan kebutuhan pengguna melalui kueri [21]. Pengambilan informasi umumnya dilakukan dari suatu kumpulan data yang beragam dan memiliki ukuran besar yang tidak efektif maupun efisien jika dilakukan pencarian manual sehingga adaptasi komputer diterapkan. *Information Retrieval System* mengimplementasikan prinsip-prinsip dan teknik temu kembali informasi ke dalam sebuah sistem untuk memudahkan pencarian informasi dari dalam suatu koleksi data seperti basis data relasional [22].

*Information Retrieval* modern umumnya melibatkan 2 tahap utama dalam memberikan informasi relevan terhadap suatu kueri yaitu tahap pengambilan (*retrieval)* dan tahap pemeringkatan (*ranking*) [23]. Tahap pengambilan bertujuan untuk mengambil semua dokumen yang relevan dengan kueri dari dalam koleksi keseluruhan dokumen tanpa mempertimbangkan urutan. Sedangkan tahap pemeringkatan mengolah dokumen relevan hasil dari tahap pengambilan untuk mencari skor relevansi tiap dokumen kemudian mengurutkannya dari skor tertinggi ke skor terendah.

## Legal Case Retrieval

Relevance judgement

## Sparse Vector Retriever

## Dense Vector Retriever

## BERT retriever.

# BAB 3

**METODOLOGI PENELITIAN**

# DAFTAR PUSTAKA

[1] Indonesia, *Undang-Undang Dasar 1945, Pasal 1 Ayat (3)*. 1945.

[2] Indonesia, Pemerintah Pusat, *Undang-Undang No. 48 Tahun 2009*. 2009.

[3] M. Y. Harahap, *Hukum Acara Perdata: Tentang Gugatan, Persidangan, Penyitaan, Pembuktian, dan Putusan Pengadilan*. Sinar Grafika, 2017.

[4] S. Mertokusumo, “Hukum acara perdata Indonesia,” 2009, Accessed: Feb. 26, 2024. [Online]. Available: http://senayan.iain-palangkaraya.ac.id/index.php?p=show\_detail&id=11461&keywords=

[5] Y. Yulia, *Hukum Acara Perdata*. Lhokseumawe: Unimal Press, 2018. Accessed: Feb. 27, 2024. [Online]. Available: https://repository.unimal.ac.id/5938/

[6] “Direktori Putusan.” Accessed: Feb. 29, 2024. [Online]. Available: https://putusan3.mahkamahagung.go.id/

[7] T. Maxwell and B. Schafer, “Natural language processing and query expansion in legal information retrieval: Challenges and a response,” *Int. Rev. Law Comput. Technol.*, vol. 24, no. 1, pp. 63–72, Mar. 2010, doi: 10.1080/13600860903570194.

[8] F. W. Lancaster and E. Gallup, “Information retrieval on-line,” 1973. Accessed: Mar. 02, 2024. [Online]. Available: https://trid.trb.org/view/9772

[9] M. S. Oghli and M. M. Almustafa, “Comparison of basic Information Retrieval Models,” *Int. J. Eng. Res. Technol.*, vol. 10, no. 09, 2021.

[10] G. Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[11] R. Baeza-Yates, “Modern Information Retrieval,” *Addison Wesley Google Sch.*, vol. 2, pp. 127–136, 1999, Accessed: Mar. 03, 2024. [Online]. Available: http://www.dcc.ufmg.br/irbook/print/chap10.ps.gz

[12] J. Ullman, “Principles of database systems Computer Science Press,” *Rockvillc Matyland*, 1980.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[14] G. G. Chowdhury, *Introduction to modern information retrieval*. Facet publishing, 2010.

[15] S. Kuzi, M. Zhang, C. Li, M. Bendersky, and M. Najork, “Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach.” arXiv, Oct. 02, 2020. doi: 10.48550/arXiv.2010.01195.

[16] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, doi: 10.1145/361219.361220.

[17] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Apr. 2009, doi: 10.1561/1500000019.

[18] N. Tonellotto, “Lecture Notes on Neural Information Retrieval.” arXiv, Sep. 12, 2022. doi: 10.48550/arXiv.2207.13443.

[19] J. Lin, R. Nogueira, and A. Yates, “Pretrained Transformers for Text Ranking: BERT and Beyond,” arXiv.org. Accessed: Mar. 06, 2024. [Online]. Available: https://arxiv.org/abs/2010.06467v3

[20] E. Q. Nuranti, E. Yulianti, and H. S. Husin, “Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents,” *Computers*, vol. 11, no. 6, Art. no. 6, Jun. 2022, doi: 10.3390/computers11060088.

[21] R. Luk, “Why is Information Retrieval a Scientific Discipline?,” *Found. Sci.*, 2020, Accessed: Mar. 07, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Why-is-Information-Retrieval-a-Scientific-Luk/0d84db991028b5da12a2aa189b374cc7baf36f86

[22] G. Kowalski, *Information Retrieval Architecture and Algorithms*. Boston, MA: Springer US, 2011. doi: 10.1007/978-1-4419-7716-8.

[23] K. A. Hambarde and H. Proenca, “Information Retrieval: Recent Advances and Beyond,” *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.