



SKRIPSI

ANALISIS PERBANDINGAN MODEL TF-IDF, *WORD2VEC*, DAN INDOBERT UNTUK SISTEM TEMU BALIK INFORMASI BERITA DENGAN TOPIK ASTA CITA

AYU LINTANG PRATIWI

NPM 22082010129

DOSEN PEMBIMBING

Abdul Rezha Efrat Najaf, S.Kom, M.Kom

Reisa Permatasari, S.T, M.Kom

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI SISTEM INFORMASI

SURABAYA

2026

LEMBAR PENGESAHAN

ANALISIS PERBANDINGAN MODEL TF-IDF, *WORD2VEC*, DAN INDOBERT UNTUK SISTEM TEMU BALIK INFORMASI BERITA DENGAN TOPIK ASTA CITA

Oleh :
AYU LINTANG PRATIWI
NPM. 22082010129

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Sistem Informasi Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada tanggal 24 Juni 2026.

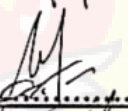
Abdul Rezha Efrat Najaf S.Kom., M.Kom
NIP. 19940929 202203 1 008


..... (Pembimbing I)

Reisa Permatasari, S.T., M.Kom
NIP. 19920514 202203 2 007


..... (Pembimbing II)

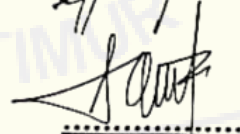
Anindo Saka Fitri, S.Kom., M.Kom
NIP. 19930325 202406 2 001


..... (Ketua Penguji)

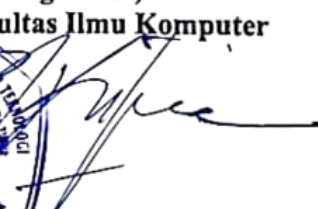
Prasasti Karunia F.A., S.Kom., M.Kom., M.IM
NIP. 19970704 202406 2 001


..... (Anggota Penguji II)

Iqbal Ramadhani Mukhlis, S.Kom., M.Kom
NIP. 19930305 202406 1 002


..... (Anggota Penguji III)

Mengetahui,
Dekan Fakultas Ilmu Komputer



Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

LEMBAR PERSETUJUAN

**ANALISIS PERBANDINGAN MODEL TF-IDF, *WORD2VEC*, DAN IN-
DOBERT UNTUK SISTEM TEMU BALIK INFORMASI BERITA
DENGAN TOPIK ASTA CITA**

Oleh :
AYU LINTANG PRATIWI
NPM. 22082010129

Menyetujui,
Koordinator Program Studi Sistem Informasi
Fakultas Ilmu Komputer



Siti Mukaromah, S.Kom, M.Kom
NIP. 19810704 2021212 011

SURAT PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini:

Nama : AYU LINTANG PRATIWI

NPM : 22082010129

Program : Sarjana (S1)

Program Studi : Sistem Informasi

Fakultas : Fakultas Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila dikemudian hari ditemukan indikasi plagiat pada Skripsi ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.



Surabaya, 13 Juli 2026
Yang Membuat Pernyataan,



AYU LINTANG PRATIWI
NPM. 22082010129

ABSTRAK

Nama Mahasiswa / NPM : Ayu Lintang Pratiwi / 22082010129
Judul Skripsi : Analisis Perbandingan Model TF-IDF, *Word2Vec*, Dan IndoBERT Untuk Sistem Temu Balik Informasi Berita dengan Topik Asta Cita
Dosen Pembimbing : 1. Abdul Rezha Efrat Najaf, S.Kom., M.Kom.
2. Reisa Permatasari, S.T, M.Kom

Peningkatan jumlah pemberitaan digital mengenai topik Asta Cita menghasilkan beragam informasi yang perlu dikelola agar pengguna dapat menemukan berita yang relevan sesuai dengan kebutuhan pencarian. Dalam sistem temu balik informasi, kualitas hasil pencarian dipengaruhi oleh metode representasi teks yang digunakan dalam merepresentasikan dokumen dan kueri. Penelitian ini bertujuan untuk membandingkan dan menganalisis performa model representasi teks TF-IDF, *Word2Vec*, dan IndoBERT dalam temu balik informasi berita bertopik Asta Cita. Tahapan dalam penelitian ini meliputi studi literatur, analisis kebutuhan, pengumpulan data, *preprocessing* teks, representasi teks masing-masing model, kemiripan dokumen dihitung dengan metode *cosine similarity*, skenario eksperimen, evaluasi performa model, analisis performa mode, dan implementasi sistem. Evaluasi dilakukan dengan beberapa skenario eksperimen dengan menggunakan metrik *precision*, *recall*, dan *Mean Average Precision* (MAP). Hasil dari skenario eksperimen menunjukkan bahwa model *Word2Vec* mampu menghasilkan dokumen yang lebih relevan dan menunjukkan hasil pemeringkatan yang lebih optimal pada korpus berita Asta Cita. Berdasarkan hasil rata-rata nilai MAP pada keseluruhan skenario, *Word2Vec* memiliki rata-rata sebesar 0,859 sedangkan TF-IDF memiliki rata-rata sebesar 0,682 dan IndoBERT sebesar 0,646. Dengan hasil tersebut, maka model *Word2Vec* yang dipilih sebagai model terbaik yang diimplementasikan dalam sistem temu balik informasi berita Asta Cita. Implementasi sistem ini menggunakan *framework Flask* guna mendukung pencarian dokumen yang sesuai dengan kebutuhan pengguna.

Kata kunci : Temu balik informasi, Asta Cita, TF-IDF, *Word2Vec*, IndoBERT

ABSTRACT

Student Name / NPM : Ayu Lintang Pratiwi / 22082010129
Thesis Title : *A Comparative Analysis of the TF-IDF, Word2Vec, and IndoBERT Models for a News Information Retrieval System on the Topic of Asta Cita*
Advisor : 1. Abdul Rezha Efrat Najaf, S.Kom., M.Kom.
2. Reisa Permatasari, S.T, M.Kom

The increase in the volume of digital news coverage on the topic of Asta Cita has generated a wide variety of information that needs to be managed so that users can find news relevant to their search needs. In information retrieval systems, the quality of search results is influenced by the text representation method used to represent documents and queries. This study aims to analyze and compare the performance of the TF-IDF, Word2Vec, and IndoBERT text representation models in retrieving news on the topic of Asta Cita. The stages of this research include a literature review, needs analysis, data collection, text preprocessing, text representation for each model, document similarity calculated using the cosine similarity method, experimental scenarios, model performance evaluation, model performance analysis, and system implementation. Evaluation was conducted using several experimental scenarios with the metrics precision, recall, and Mean Average Precision (MAP). The results of this study indicate that the Word2Vec model is capable of generating more relevant documents and achieving more optimal ranking results on the Asta Cita news corpus. Based on the average MAP scores for all scenarios, Word2Vec had an average score of 0.859, while TF-IDF had an average score of 0.682 and IndoBERT had an average score of 0.646. Based on these results, the Word2Vec model was selected as the best model to be implemented in the Asta Cita news information retrieval system. The system was implemented using the Flask framework to support document searches tailored to user needs.

Keywords: *Information retrieval, Asta Cita, TF-IDF, Word2Vec, IndoBERT*

KATA PENGANTAR

Puji syukur kehadiran Allah SWT atas segala rahmat, karunia, dan hidayah-Nya, sehingga penulis dapat menyelesaikan penyusunan dan penelitian skripsi yang berjudul “Analisis Perbandingan Model TF-IDF, *Word2Vec*, Dan IndoBERT Untuk Sistem Temu Balik Informasi Berita Dengan Topik Asta Cita” dengan baik. Skripsi ini disusun sebagai salah satu persyaratan untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur.

Dalam proses penyusunan skripsi ini, penulis memperoleh banyak bantuan, arahan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Orang tua dan keluarga yang senantiasa mendoakan dan memberikan dukungan moril dan materiil selama proses perkuliahan dan penyusunan skripsi.
2. Ibu Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur atas dukungan dan fasilitas akademik yang diberikan selama masa perkuliahan.
3. Ibu Siti Mukaromah, S.Kom., M.Kom., selaku Kepala Program Studi Sistem Informasi atas dukungan serta fasilitas akademik selama proses perkuliahan.
4. Bapak Abdul Rezha Efrat Najaf, S.Kom., M.Kom., selaku dosen Pembimbing I yang telah memberikan bimbingan, arahan, motivasi, serta saran dalam proses penelitian dan penyusunan skripsi.
5. Ibu Reisa Permatasari, S.T., M.Kom., selaku dosen Pembimbing II yang telah memberikan bimbingan, arahan, motivasi, serta saran dalam proses penelitian dan penyusunan skripsi. Seluruh dosen dan staf Program Studi Sistem Informasi atas segala ilmu dan bantuan yang diberikan selama proses perkuliahan.
6. Teman-teman terdekat saya yaitu, Agnes Bella Trisanda, Sadrah Zefanya Putra, Putri Nurul Hidayati, Ramdhan Ariansyah, Aisha Nurliana, dan Dea Ananda Refiza yang senantiasa memberikan dukungan dan semangat kepada penulis dalam masa perkuliahan.

7. Teman-teman dan Bapak/Ibu dosen yang berada di LSP yang telah memberikan dukungan, semangat, bantuan, serta kebersamaan kepada penulis selama proses perkuliahan dan penyusunan skripsi.
8. Teman-teman yang selalu hadir memberikan bantuan, semangat, serta kebersamaan selama masa perkuliahan hingga penyusunan skripsi.
9. Serta semua pihak yang telah membantu dan terlibat dalam proses penyusunan proposal ini, yang tidak dapat penulis sebutkan satu per satu.
10. Kopi Kenangan Merr Rungkut Surabaya yang telah memberikan tempat, suasana, serta dukungan kepada penulis selama proses pengerjaan dan penyusunan skripsi.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan guna penyempurnaan di masa mendatang. Penulis berharap skripsi ini dapat memberikan manfaat, menambah wawasan, serta menjadi referensi yang berguna bagi pembaca maupun penelitian selanjutnya.

Surabaya, 10 Juni 2026

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN	Error! Bookmark not defined.
SURAT PERNYATAAN ORISINALITAS	Error! Bookmark not defined.
ABSTRAK	ix
KATA PENGANTAR.....	xiii
DAFTAR ISI.....	xv
DAFTAR GAMBAR.....	xix
DAFTAR TABEL	xxi
DAFTAR KODE	xxiii
DAFTAR LAMPIRAN	xxv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	4
1.3. Batasan Masalah.....	5
1.4. Tujuan Penelitian	5
1.5. Manfaat Penelitian	6
1.6. Sistematika Penulisan	6
BAB II TINJAUAN PUSTAKA.....	9
2.1. Dasar Teori.....	9
2.1.1. <i>Text Mining</i>	9
2.1.2. <i>Information Retrieval (IR)</i>	9
2.1.3 Representasi Teks.....	10
2.1.4 <i>Natural Language Processing (NLP)</i>	11
2.1.5 <i>Word Embedding</i>	11
2.1.6 <i>Web Scraping</i>	12

2.1.7	Asta Cita	12
2.1.8	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	13
2.1.9	<i>Word2Vec</i>	14
2.1.10	IndoBERT	16
2.1.11	<i>Cosine Similarity</i>	16
2.1.12	<i>Precision</i>	17
2.1.13	<i>Recall</i>	18
2.1.14	<i>Mean Average Precision (MAP)</i>	19
2.1.15	<i>Flask</i>	19
2.2.	Penelitian Terdahulu	20
BAB III METODOLOGI PENELITIAN		25
3.1.	Studi Literatur	26
3.2.	Analisis Kebutuhan	26
3.3.	Pengumpulan Data	27
3.4.	<i>Preprocessing</i> Teks	28
3.4.1	<i>Cleaning</i>	29
3.4.2	<i>Case Folding</i>	29
3.4.3	Tokenisasi	29
3.4.4	<i>Stopword Removal</i>	29
3.4.5	<i>Stemming</i>	30
3.5.	Representasi Teks	30
3.6.	Perhitungan <i>Similarity</i> & Ranking	33
3.7.	Skenario Eksperimen	33
3.8.	Evaluasi Kinerja	35
3.9.	Implementasi dan <i>Deployment</i> Sistem	35
BAB IV HASIL DAN PEMBAHASAN		37

4.1.	Studi Literatur	37
4.2.	Analisis Kebutuhan	37
4.2.1	Kebutuhan Data.....	37
4.2.2	Kebutuhan Perangkat Keras dan Lunak.....	38
4.3.	Pengumpulan Data	38
4.4.	<i>Preprocessing</i> Teks.....	41
4.4.1	<i>Cleaning</i>	42
4.4.2	Case Folding	48
4.4.3	Tokenisasi	49
4.4.4	Stopword Removal.....	51
4.4.5	Stemming	53
4.5.	Representasi Teks.....	54
4.5.1.	Representasi Teks dengan TF-IDF.....	55
4.5.2.	Representasi Teks dengan <i>Word2Vec</i>	56
4.5.3.	Representasi Teks dengan IndoBERT	59
4.6.	Perhitungan Similarity dan Ranking	61
4.6.1	Perhitungan Similarity dan Ranking dengan TF-IDF	61
4.6.2	Perhitungan Similarity dan Ranking dengan <i>Word2Vec</i>	63
4.6.3	Perhitungan Similarity dan Ranking dengan IndoBERT	65
4.7.	Skenario Eksperimen	67
4.7.1	Skenario Perbandingan Antar Model	67
4.7.2	Skenario Variasi Jenis Kueri.....	73
4.7.3	Skenario Variasi Panjang Kueri	81
4.7.4	Skenario Variasi Jumlah Hasil Pencarian	91
4.7.5	Skenario Variasi Ruang Pencarian.....	97
4.8.	Evaluasi Kinerja	102

4.9.	Implementasi dan <i>Deployment</i> Sistem	106
4.9.1	Wireframe	106
4.9.2	Implementasi Sistem	110
BAB V PENUTUP		117
5.1	Kesimpulan.....	117
5.2	Saran	118
DAFTAR PUSTAKA		119
LAMPIRAN		123

DAFTAR GAMBAR

Gambar 2.1 Arsitektur Skip-gram.....	15
Gambar 2.2 Grafik Cosine Similarity	16
Gambar 3.1 Alur Metodologi Penelitian.....	25
Gambar 3.2 Alur TF-IDF	31
Gambar 3.3 Alur <i>Word2Vec</i>	32
Gambar 3.4 Alur IndoBERT	32
Gambar 4. 1 Kata Yang Sering Muncul.....	41
Gambar 4. 2 Ukuran Matrik TF-IDF.....	56
Gambar 4. 3 Grafik Nilai MAP	104
Gambar 4. 4 Rancangan Wireframe Halaman Home.....	107
Gambar 4. 6 Rancangan Wireframe Halaman Hasil Pencarian	108
Gambar 4. 7 Rancangan Wireframe Halaman Detail Berita.....	109
Gambar 4. 8 Rancangan Wireframe Halaman Insight berita	110
Gambar 4. 9 Implementasi Sistem Halaman Home	111
Gambar 4. 10 Implementasi Sistem Halaman Pencarian	112
Gambar 4. 11 Implementasi Sistem Halaman Hasil Pencarian 1.....	113
Gambar 4. 12 Implementasi Sistem Halaman Hasil Pencarian 2.....	113
Gambar 4. 13 Implementasi Sistem Halaman Hasil Pencarian 3.....	113
Gambar 4. 14 Implementasi Sistem Halaman Detail Berita	114
Gambar 4. 15 Implementasi Sistem Halaman <i>Insight</i> berita 1.....	115
Gambar 4. 16 Implementasi Sistem Halaman <i>Insight</i> berita 2.....	115
Gambar 4. 17 Implementasi Sistem Halaman <i>Insight</i> Berita 3	116

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu	20
Tabel 3.1 Keyword Pengumpulan Data	27
Tabel 4. 1 Hasil Scraping Berita	39
Tabel 4. 2 Hasil Data Cleaning	46
Tabel 4. 3 Hasil Case Folding	48
Tabel 4. 4 Hasil Tokenisasi	50
Tabel 4. 5 Hasil Stopword Removal	52
Tabel 4. 6 Hasil Stemming	54
Tabel 4. 7 Hasil Pembobotan TF-IDF	56
Tabel 4. 8 Hasil Vektor <i>Word2Vec</i>	58
Tabel 4. 9 Hasil Vektor IndoBERT	60
Tabel 4. 10 Hasil Perankingan dan Similarity Model TF-IDF	62
Tabel 4. 11 Hasil Perankingan dan Similarity Model <i>Word2Vec</i>	64
Tabel 4. 12 Hasil Perankingan dan Similarity Model IndoBERT	66
Tabel 4. 13 Kueri Skenario Perbandingan Antar Model	67
Tabel 4. 14 Hasil Dokumen Kueri 1 pada Skenario Pertama	68
Tabel 4. 15 Hasil Dokumen Kueri 2 pada Skenario Pertama	69
Tabel 4. 16 Hasil Dokumen Kueri 3 pada Skenario Pertama	70
Tabel 4. 17 Ringkasan Hasil Skenario Perbandingan Antar Model	71
Tabel 4. 18 Contoh Skenario Variasi Jenis Kueri	73
Tabel 4. 19 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Keyword Asta Cita	73
Tabel 4. 20 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Semantik Terkait	74
Tabel 4. 21 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Gabungan Konsep	75

Tabel 4. 22 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Keyword Asta Cita	76
Tabel 4. 23 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Semantik Terkait	77
Tabel 4. 24 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Gabungan Konsep	78
Tabel 4. 25 Ringkasan Hasil Skenario Variasi Jenis Kueri.....	79
Tabel 4. 26 Contoh Kueri Skenario Panjang Kueri	82
Tabel 4. 27 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Pendek.....	82
Tabel 4. 28 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Sedang.....	83
Tabel 4. 29 Hasil Retrieval Dokumen Kueri 1 pada Jenis Kueri Panjang.....	84
Tabel 4. 30 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Pendek.....	85
Tabel 4. 31 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Sedang.....	86
Tabel 4. 32 Hasil Retrieval Dokumen Kueri 2 pada Jenis Kueri Panjang.....	87
Tabel 4. 33 Ringkasan Hasil Skenario Variasi Panjang Kueri	88
Tabel 4. 34 Hasil Retrieval Dokumen Pada Pencarian Top-10.....	91
Tabel 4. 35 Hasil Retrieval Dokumen Pada Pencarian Top-20.....	93
Tabel 4. 36 Hasil Retrieval Dokumen Pada Pencarian Top-50.....	94
Tabel 4. 37 Ringkasan Hasil Skenario Variasi Jumlah Pencarian.....	96
Tabel 4. 38 Hasil Retrieval Dokumen Pada Pencarian Dengan Korpus 100%	97
Tabel 4. 39 Hasil Retrieval Dokumen Pada Pencarian Dengan Korpus 75%	98
Tabel 4. 40 Hasil Retrieval Dokumen Pada Pencarian Dengan Korpus 50%	99
Tabel 4. 41 Ringkasan Hasil Skenario Variasi Ruang Pencarian.....	100
Tabel 4. 42 Hasil Evaluasi Kinerja.....	102
Tabel 4. 43 Rata-Rata Nilai MAP Seluruh Skenario.....	104

DAFTAR KODE

Kode 4. 1 Pengumpulan Data	39
Kode 4. 2 Menghapus Angka.....	42
Kode 4. 3 Kode Menghapus Awalan Berita.....	43
Kode 4. 4 Kode Menghapus Tanggal dan Jam	43
Kode 4. 5 Kode Menghapus Footer Berita.....	44
Kode 4. 6 Kode Menghapus Spasi Berlebih	44
Kode 4. 7 Kode Menghapus Simbol	44
Kode 4. 8 Kode Menghapus Konten Non-Relevan.....	45
Kode 4. 9 Kode Menghapus Elemen Tambahan Berita	45
Kode 4. 10 Kode Case Folding	48
Kode 4. 11 Kode Tokenisasi	50
Kode 4. 12 Kode Stopword Removal	52
Kode 4. 13 Kode Stemming	53
Kode 4. 14 Pembuatan Vektor TF-IDF	55
Kode 4. 15 Pelatihan Model <i>Word2Vec</i>	57
Kode 4. 16 Pembentukan Vektor <i>Word2Vec</i>	57
Kode 4. 17 Inisialisasi Model dan Tokenizer IndoBERT Pretrained.....	59
Kode 4. 18 Pembentukan Vektor IndoBERT	60
Kode 4. 19 Perhitungan Similarity Model TF-IDF	62
Kode 4. 20 Perhitungan Similarity Model <i>Word2Vec</i>	64
Kode 4. 21 Perhitungan Similarity Model IndoBERT	66

DAFTAR LAMPIRAN

Lampiran 1. Data Scraping Asta Cita	123
Lampiran 2. Contoh Dokumen Hasil Retrieval.....	124