

**PERBANDINGAN SISTEM TEMU BALIK METODE BM25
DAN METODE NEURAL NETWORK DENGAN BERT
PREPROCESSING PADA DATASET CRANFIELD**

TUGAS AKHIR

**Andreas Aditya
1118008**



**PROGRAM STUDI INFORMATIKA
INSTITUT TEKNOLOGI HARAPAN BANGSA
BANDUNG
2022**

**PERBANDINGAN SISTEM TEMU BALIK METODE BM25
DAN METODE NEURAL NETWORK DENGAN BERT
PREPROCESSING PADA DATASET CRANFIELD**

TUGAS AKHIR

**Diajukan sebagai salah satu syarat untuk memperoleh
gelar sarjana dalam bidang Informatika**

Andreas Aditya

1118008



INSTITUT
TEKNOLOGI
HARAPAN
BANGSA

Veritas vos liberabit

**PROGRAM STUDI INFORMATIKA
INSTITUT TEKNOLOGI HARAPAN BANGSA
BANDUNG
2022**

ABSTRAK

Nama : Andreas Aditya
Program Studi : Informatika
Judul : Perbandingan Sistem Temu Balik Metode BM25 dan Metode Neural Network dengan BERT Preprocessing pada Dataset Cranfield

Perkembangan sistem temu balik informasi memunculkan berbagai metode *machine learning* yang menggunakan konteks dari dokumen untuk menemukan informasi yang dicari. Model BERT yang digunakan oleh *Google* merupakan model terbaik dari sistem pencarian situs di internet. Penelitian ini membandingkan BERT yang digunakan sebagai komponen *preprocessing* dokumen dan *query* untuk model *neural network* dengan metode BM25 pada *dataset Cranfield*. Metode BM25 memiliki 2 indikator yaitu *b* dan *k* yang masing-masing berfokus pada *weight* untuk panjang dokumen dan frekuensi kemunculan kata. Indikator yang digunakan untuk metode *neural network* dengan BERT *preprocessing* adalah ukuran *subsegment* sebuah dokumen, *learning rate*, ukuran *layer neural network*, dan metode penggabungan nilai relevansi *subsegment*. Dataset yang digunakan dalam penelitian ini adalah dataset Cranfield yang berisi 1400 dokumen dan 225 *query*. Berdasarkan pengujian semua indikator, metode BM25 memberikan nilai *nDCG@20* sebesar 0.5054 dan metode *neural network* dengan BERT *preprocessing* menghasilkan nilai *nDCG@20* sebesar 0.5520.

Kata kunci: *Information Retrieval, Deep Learning, BM25, BERT, Natural Language Processing (NLP)*.

ABSTRACT

Name : Nama Pengarang
Department : Informatics
Title : Information Retrieval System Comparison between
BM25 Method and Neural Network Method with BERT
Preprocessing on Cranfield Dataset

The development of information retrieval systems has led to various machine learning methods that able to learn document's context to find the relevant information. BERT model used by Google currently is a great model for information retrieval system on the internet. This study compares neural network model with BERT preprocessing and the BM25 method on Cranfield dataset. The BM25 method has 2 indicators, namely b and k , each giving weight for the length of the document and the frequency of occurrence of words. The indicators used for the neural network method with BERT preprocessing are the size of the document subsegment, learning rate, neural network layer size, and the method of combining relevance values from subsegments. The dataset used in this research is the Cranfield dataset which contains 1400 documents and 225 queries. Based on the results, the BM25 method gives an $nDCG@20$ value of 0.5054 and the neural network method with BERT preprocessing produces an $nDCG@20$ value of 0.5520.

Keywords: Information Retrieval, Deep Learning, BM25, BERT, Natural Language Processing (NLP).

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan yang Maha Esa karena atas karunia dan bimbingan-Nya penulis berhasil menyelesaikan tugas akhir yang berjudul "Perbandingan Sistem Temu Balik Metode BM25 dan Metode Neural Network dengan BERT Preprocessing pada Dataset Cranfield". Laporan ini disusun sebagai salah satu syarat kelulusan di Institut Teknologi Harapan Bangsa. Pada kesempatan ini penulis menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Tuhan Yesus dan Bunda Maria, karena bimbingan dan karunia-Nya, penulis selalu mendapat pengharapan, kebahagiaan, serta pengalaman berharga selama pembuatan Tugas Akhir ini.
2. Ibu Oviliani Yenty Yuliana, S.T., MSCIS, Ph.D., selaku Pembimbing I Tugas Akhir yang senantiasa memberi semangat, ilmu-ilmu, saran, dan dukungan kepada penulis selama tugas akhir berlangsung dan selama pembuatan laporan tugas akhir ini.
3. Bapak Dr. Hery Heryanto, M.Kom., selaku pembimbing II tugas akhir yang senantiasa memberi semangat, ilmu-ilmu, saran, dan dukungan kepada penulis selama tugas akhir berlangsung dan selama pembuatan laporan tugas akhir ini.
4. Bapak Ventje Jeremias Lewi Engel, M.T., CEH, selaku Penguji I tugas akhir. Terima kasih atas semangat, ilmu-ilmu, saran, dan dukungan kepada penulis selama tugas akhir berlangsung dan selama pembuatan laporan tugas akhir ini.
5. Seluruh dosen dan staff Departemen Teknik Informatika ITHB yang telah membantu dalam menyelesaikan Laporan Tugas Akhir ini.
6. Segenap jajaran staf dan karyawan ITHB yang turut membantu kelancaran dalam menyelesaikan Laporan Tugas Akhir ini.
7. Kedua orang tua yang menyediakan waktu untuk memberikan doa, semangat, dan dukungan kepada penulis serta sabar menunggu dalam menyelesaikan tugas akhir ini.
8. Hanjaya Suryalim, Daniel Christianto, Alexander Tri Handoyo, Katherine

Rayani, Benedict Reydo Bayuputra, Daniel Alexander, dan Gabriela Solaiman yang sudah mendukung dan memberikan semangat serta berjuang bersama untuk menyelesaikan tugas akhir.

Penulis menyadari bahwa laporan ini masih jauh dari sempurna karena keterbatasan waktu dan pengetahuan yang dimiliki oleh penulis. Oleh karena itu, kritik dan saran untuk membangun kesempurnaan tugas akhir ini sangat diharapkan. Semoga dengan adanya tugas akhir ini membantu pihak yang membutuhkan.

Akhir kata, semoga tugas akhir ini membantu pembaca yang hendak mengetahui teknologi yang digunakan untuk sistem temu balik informasi.

Bandung, 1 Juni 2022

Hormat penulis,

A handwritten signature in black ink, appearing to read 'Andreas Aditya', with a stylized flourish at the end.

Andreas Aditya

DAFTAR ISI

ABSTRAK	iv
ABSTRACT	v
KATA PENGANTAR	vi
DAFTAR ISI	viii
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiv
DAFTAR ALGORITMA	xv
DAFTAR LAMPIRAN	xvi
BAB 1 PENDAHULUAN	1-1
1.1 Latar Belakang	1-1
1.2 Rumusan Masalah	1-2
1.3 Tujuan Penelitian	1-2
1.4 Batasan Masalah	1-3
1.5 Kontribusi Penelitian	1-3
1.6 Metodologi Penelitian	1-3
1.7 Sistematika Pembahasan	1-4
BAB 2 LANDASAN TEORI	2-1
2.1 Tinjauan Pustaka	2-1
2.1.1 Sistem Temu Balik Informasi	2-1
2.1.1.1 Pemberian Peringkat	2-2
2.1.1.2 Penilaian Sistem Temu Balik Informasi	2-2
2.1.2 <i>Natural Language Processing</i>	2-3
2.1.3 BM25	2-4
2.1.4 <i>Deep Learning</i>	2-5
2.1.5 <i>Transformer</i>	2-6
2.1.5.1 <i>Encoder Block</i>	2-7
2.1.5.2 <i>Decoder Block</i>	2-8
2.1.5.3 <i>Multi-Head Attention</i>	2-9

2.1.5.4	<i>Scaled Dot-Product Attention</i>	2-10
2.1.6	BERT	2-11
2.2	Pustaka <i>Python</i>	2-12
2.2.1	<i>Numpy</i>	2-13
2.2.2	<i>Scikit Learn</i>	2-13
2.2.3	<i>Tensorflow</i>	2-14
2.2.4	<i>Natural Language Toolkit</i>	2-16
2.2.5	<i>Flask</i>	2-17
2.3	Tinjauan Studi	2-18
2.4	Tinjauan Objek	2-23
BAB 3	ANALISIS DAN PERANCANGAN SISTEM	3-1
3.1	Analisis Masalah	3-1
3.2	Kerangka Pemikiran	3-1
3.3	Urutan Proses <i>Global</i>	3-4
3.3.1	Proses <i>Global</i> Metode BM25	3-4
3.3.2	Proses <i>Global</i> Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	3-5
3.4	Analisis Manual	3-6
3.4.1	Proses Ranking dengan Metode BM25	3-6
3.4.2	<i>Preprocess</i> Dokumen dan <i>Query</i> dengan BERT	3-9
3.4.3	Arsitektur Model	3-15
3.4.4	Evaluasi Model	3-15
BAB 4	IMPLEMENTASI DAN PENGUJIAN	4-1
4.1	Lingkungan Implementasi	4-1
4.1.1	Spesifikasi Perangkat Keras	4-1
4.1.2	Spesifikasi Perangkat Lunak	4-1
4.2	Implementasi Perangkat Lunak	4-1
4.2.1	Daftar <i>Class</i> dan <i>Method</i>	4-2
4.2.1.1	<i>Class</i> BM25	4-2
4.2.1.2	<i>Class BERT_preprocessing</i>	4-3
4.2.1.3	<i>Class</i> train_model	4-4
4.2.1.4	<i>Class</i> test_model	4-5
4.2.2	Implementasi BM25	4-6
4.2.3	Implementasi Metode <i>Neural Network</i> dengan Bert <i>Preprocessing</i>	4-10
4.2.4	Implementasi Aplikasi	4-11

4.3	Pengujian	4-12
4.3.1	Pengujian Parameter BM25	4-12
4.3.2	Pengujian Ukuran <i>Subsegment</i> untuk Model BERT	4-13
4.3.3	Pengujian Ukuran <i>Layer</i> untuk Metode <i>Neural Network</i>	4-13
4.3.4	Pengujian Nilai <i>Learning Rate</i> untuk Metode <i>Neural Network</i>	4-13
4.3.5	Pengujian Metode Penggabungan Nilai Relevansi Dokumen untuk Metode <i>Neural Network</i>	4-14
4.4	Hasil Pengujian	4-14
4.4.1	Hasil Pengujian BM25	4-15
4.4.2	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>preprocessing</i>	4-16
4.4.2.1	Analisa Pengujian Metode <i>Neural Network</i> dengan BERT <i>preprocessing</i> Berdasarkan Ukuran <i>Subsegment</i>	4-17
4.4.2.2	Analisa Pengujian Metode <i>Neural Network</i> dengan BERT <i>preprocessing</i> Berdasarkan Ukuran <i>Layer</i>	4-18
4.4.2.3	Analisa Pengujian Metode <i>Neural Network</i> dengan BERT <i>preprocessing</i> Berdasarkan <i>Learning Rate</i>	4-19
4.4.2.4	Analisa Pengujian Metode <i>Neural Network</i> dengan BERT <i>preprocessing</i> Berdasarkan Metode Penggabungan Nilai Relevansi	4-20
4.4.2.5	Analisa Keseluruhan Pengujian Metode <i>Neural</i> <i>Network</i> dengan BERT <i>preprocessing</i>	4-22
4.4.3	Waktu Pengujian	4-26
4.4.3.1	Waktu Pengujian Metode BM25	4-26
4.4.3.2	Waktu Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	4-27
4.4.3.3	Waktu Pengujian Berdasarkan Jumlah Query	4-28
4.4.4	Membandingkan Hasil Pengujian Terbaik dari Kedua Metode dengan Penelitian Sebelumnya	4-29
4.5	Analisis Kesalahan	4-30
BAB 5	KESIMPULAN DAN SARAN	5-1
5.1	Kesimpulan	5-1
5.2	Saran	5-2

BAB A Hasil Pengujian Metode BM25	A-1
BAB B Hasil Pengujian Metode Neural Network dengan BERT Preprocessing	B-5
BAB C Waktu Pencarian Metode BM25	C-20
BAB D Waktu Pencarian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	D-24
BAB E Waktu Pengujian Metode BM25 Berdasarkan Jumlah <i>Query</i>	E-27
BAB F Waktu Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Jumlah <i>Query</i>	F-29

DAFTAR TABEL

2.1	Daftar <i>Method</i> yang Digunakan dari Pustaka <i>Numpy</i>	2-13
2.2	Daftar <i>Method</i> yang Digunakan dari Pustaka <i>Scikit Learn</i>	2-14
2.3	Daftar <i>Method</i> yang Digunakan dari Pustaka <i>tensorflow</i>	2-15
2.4	Daftar <i>Method</i> yang Digunakan dari Pustaka <i>NLTK</i>	2-17
2.5	Daftar <i>Method</i> yang Digunakan dari Pustaka <i>Flask</i>	2-17
2.6	Tinjauan Studi	2-18
3.1	Lima Dokumen dengan Nilai Relevansi Tertinggi dari Hasil Metode BM25	3-8
3.2	Label Relevansi Dokumen	3-8
3.3	Contoh Perhitungan nDCG	3-9
3.4	Contoh Kata-kata dan <i>token</i> yang Didapat	3-13
3.5	<i>Query tokenization</i>	3-13
3.6	Contoh Nilai Relevansi Dokumen	3-16
3.7	Proses Mengurutkan Dokumen	3-16
3.8	Label Relevansi Dokumen	3-17
3.9	Contoh perhitungan nDCG	3-17
4.1	Daftar <i>Method</i> pada Class BM25	4-2
4.2	Daftar <i>Method</i> pada Class <i>BERT_preprocessing</i>	4-3
4.3	Daftar <i>Method</i> pada Class <i>train_model</i>	4-4
4.4	Daftar <i>Method</i> pada Class <i>test_model</i>	4-5
4.5	Statistik Deskriptif Hasil Pengujian Metode <i>Neural Network</i> dengan <i>BERT Preprocessing</i> Berdasarkan Ukuran <i>Subsegment</i>	4-18
4.6	Statistik Deskriptif Hasil Pengujian Metode <i>Neural Network</i> dengan <i>BERT Preprocessing</i> Berdasarkan Ukuran <i>Layer</i>	4-19
4.7	Statistik Deskriptif Hasil Pengujian Metode <i>Neural Network</i> dengan <i>BERT Preprocessing</i> Berdasarkan <i>Learning Rate</i>	4-20
4.8	Statistik Deskriptif Hasil Pengujian Metode <i>Neural Network</i> dengan <i>BERT Preprocessing</i> Berdasarkan Metode Penggabungan Nilai Relevansi.	4-21
4.9	Sepuluh Hasil Pengujian Terbaik dari Metode <i>Neural Network</i> dengan <i>BERT Preprocessing</i>	4-25
4.10	Waktu Pengujian Metode BM25	4-27

4.11 Waktu Pengujian Metode Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	4-27
4.12 Hasil Pengujian Terbaik dari Penelitian Ini dan Peneklitian Negin et al. [6]	4-29
A-1 Hasil Pengujian Metode BM25	A-1
B-1 Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	B-5
C-1 Waktu Pencarian Metode BM25	C-20
D-1 Waktu Perhitungan Nilai Relevansi Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	D-24
E-1 Waktu Pengujian Metode BM25 Berdasarkan Jumlah <i>Query</i>	E-27
F-1 Waktu Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Jumlah <i>Query</i>	F-29

DAFTAR GAMBAR

2.1	Proses Pemberian Peringkat Pada Dokumen	2-2
2.2	Arsitektur <i>Transformer</i>	2-7
2.3	Blok <i>Encoder</i> dari Arsitektur <i>Transformer</i>	2-7
2.4	Blok <i>Decoder</i> dari Arsitektur <i>Transformer</i>	2-8
2.5	<i>Multi-Head Attention</i> dari Arsitektur <i>Transformer</i>	2-10
2.6	<i>Scaled Dot-Product Attention</i> dari Arsitektur <i>Transformer</i>	2-10
2.7	Arsitektur BERT	2-12
2.8	BERT <i>Input</i>	2-12
2.9	Contoh Dokumen	2-24
2.10	Contoh <i>Query</i>	2-24
2.11	Contoh Relasi antara <i>Query</i> dan Dokumen	2-25
3.1	Kerangka Pemikiran	3-1
3.2	Urutan Proses Global	3-4
3.3	Contoh Dokumen	3-6
3.4	Contoh dokumen yang telah dibersihkan	3-7
3.5	Informasi dari dokumen	3-7
3.6	Contoh <i>Query</i>	3-7
3.7	<i>Query</i> yang telah dibersihkan	3-8
3.8	Contoh Dokumen	3-10
3.9	<i>tokenized Document</i>	3-11
3.10	<i>Subsegments</i>	3-12
3.11	<i>tokenized Subsegment</i>	3-12
3.12	<i>tokenized Query Subsegment Pair</i>	3-14
3.13	<i>Neural Network Input</i>	3-14
3.14	Arsitektur Model	3-15
4.1	Tampilan Penggunaan Metode BM25	4-11
4.2	Tampilan Penggunaan Metode Neural Network dengan BERT Preprocessing	4-12
4.3	Hasil Pencarian	4-12
4.4	<i>Heatmap</i> Hasil Pengujian BM25	4-16
4.5	Contoh <i>Boxplot</i>	4-17
4.6	<i>Boxplot</i> Hasil Pengujian Metode Neural Network dengan BERT Preprocessing Berdasarkan Ukuran <i>Subsegment</i>	4-18

4.7	<i>Boxplot</i> Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Ukuran <i>Layer</i>	4-19
4.8	<i>Boxplot</i> Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan <i>Learning Rate</i>	4-20
4.9	<i>Boxplot</i> Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Metode Penggabungan Nilai Relevansi .	4-21
4.10	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	4-22
4.11	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> berdasarkan Ukuran <i>Subsegment</i>	4-22
4.12	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> berdasarkan Metode Penggabungan Nilai Relevansi .	4-23
4.13	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> berdasarkan Ukuran <i>Layer</i>	4-24
4.14	Hasil Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> yang Diurutkan Ulang	4-24
4.15	Waktu Pengujian Metode BM25	4-26
4.16	Waktu Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i>	4-27
4.17	Waktu Pengujian Metode BM25 Berdasarkan Jumlah <i>Query</i>	4-28
4.18	Waktu Pengujian Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Jumlah <i>Query</i>	4-28
4.19	Waktu Pengujian Metode BM25 dan Metode <i>Neural Network</i> dengan BERT <i>Preprocessing</i> Berdasarkan Jumlah <i>Query</i>	4-29

DAFTAR ALGORITMA

4.1	Preprocess Document	4-7
4.2	Calculate Documents Attribute	4-8
4.3	BM25	4-9
4.4	Calculate nDCG@20	4-10

DAFTAR LAMPIRAN

LAMPIRAN A	A-1
LAMPIRAN B	B-5
LAMPIRAN C	C-20
LAMPIRAN D	D-24
LAMPIRAN E	E-27
LAMPIRAN F	F-29

DAFTAR REFERENSI

- [1] Dai, Zhuyun, and Jamie Callan. "Deeper text understanding for IR with contextual neural language modeling." *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019.
- [2] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- [3] Robertson, Stephen, and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc, 2009.
- [4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [5] Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge: Cambridge University Press, 2008.
- [6] Ghasemi, Negin, and Djoerd Hiemstra. "BERT meets Cranfield: Uncovering the Properties of Full Ranking on Fully Labeled Data." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2021.
- [7] Hapke, Hannes, Cole Howard, and Hobson Lane. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Simon and Schuster, 2019.
- [8] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
- [9] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
- [10] Harris, Charles R., et al. "Array programming with NumPy." *Nature* 585.7825 (2020): 357-362.
- [11] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

DAFTAR REFERENSI

- [12] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.
- [13] Han, Shuguang, et al. "Learning-to-rank with bert in tf-ranking." *arXiv preprint arXiv:2004.08476* (2020).
- [14] Nogueira, Rodrigo, and Kyunghyun Cho. "*Passage Re-ranking with BERT.*" *arXiv preprint arXiv:1901.04085* (2019).
- [15] Heaton, Jeff. *Introduction to neural networks with Java*. Heaton Research, Inc., 2008.