

# Aplikasi *Bidirectional Encoder Representations from Transformers* untuk Pemeringkatan Teks Bahasa Indonesia

Carles Octavianus

Dosen Pembimbing: Sarini Abdullah S.Si., M.Stats., Ph.D.

3 Januari, 2024



# Daftar Isi

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT



# Table of Contents

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT



# Pendahuluan

1. Peningkatan jumlah data teks digital membuat manusia kesulitan dalam memproses informasi secara efektif dan efisien.
2. Tahap pertama dalam memproses informasi dalam data teks adalah melakukan penyimpanan data teks dengan efisien.
3. Diperlukan mekanisme untuk mengembalikan teks yang relevan dari data teks dan mekanisme pengembalian informasi menjadi semakin penting dengan peningkatan jumlah data teks.



# Pendahuluan

1. Pemeringkatan teks adalah salah satu mekanisme untuk mengembalikan teks yang relevan.
2. Tujuan dari pemeringkatan teks adalah menghasilkan daftar teks yang terurut berdasarkan relevansinya terhadap permintaan pengguna.
3. Pemeringkatan teks banyak digunakan dalam mesin pencarian untuk menghasilkan daftar teks yang relevan.



# Contoh Pemeringkatan Teks

a3 size


AI Images Maps News Videos More Settings Tools

About 289,000,000 results (0.73 seconds)

**21.0 x 29.7cm**

The **A3** size print measures 29.7 x 42.0cm, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4** size print measures 21.0 x 29.7cm, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)  
[https://www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)



About Featured Snippets Feedback

Google.com – 14.10.2019

a3 size


AI Images Maps News Videos More Settings Tools

About 846,000,000 results (0.50 seconds)

**29.7 x 42.0cm**

The **A3** size print measures **29.7 x 42.0cm**, 11.69 x 16.53 inches, if mounted 40.6 x 50.8cm, 15.98 x 20 inches. The **A4** size print measures **21.0 x 29.7cm**, 8.27 x 11.69 inches, if mounted 30.3 x 40.6cm, 11.93 x 15.98 inches.

[www.stephenwiltshire.co.uk/paper\\_sizes](https://www.stephenwiltshire.co.uk/paper_sizes)  
[Paper Sizes A0, A1, A2, A3, A4 - Stephen Wiltshire](https://www.stephenwiltshire.co.uk/paper_sizes)

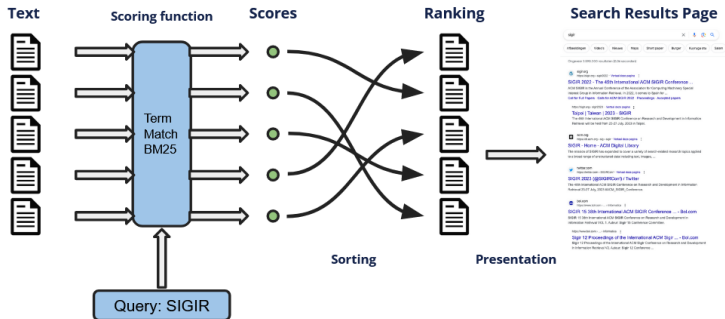


About Featured Snippets Feedback

Google.com – 3.3.2020



# Classical Text Ranking



## Vocabulary Mismatch

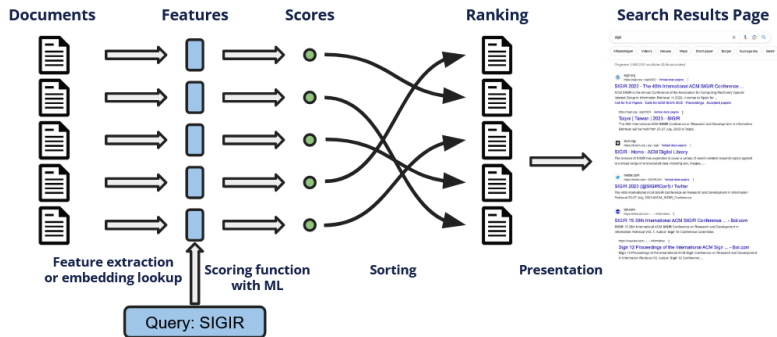
1. Pemeringkatan teks dengan kecocokan kata antara kueri dan teks memiliki kelemahan, yaitu sistem tidak dapat mengambil teks yang relevan bila kueri dan teks memiliki kata yang berbeda.
2. Sebagai contoh, untuk kueri apa makanan ternenak di Indonesia, teks dengan kalimat hidangan terlezat di nusantara adalah rendang tentunya akan mendapatkan skor yang rendah karena tidak memiliki kata yang sama dengan kueri.
3. Dengan memanfaatkan data tambahan seperti log atau jumlah klik (pada web) dari teks, model *machine learning* dapat digunakan untuk memeringkatkan teks.





# Alur Pemeringkatan Teks dengan *Machine Learning*

## Text Ranking with Classical ML



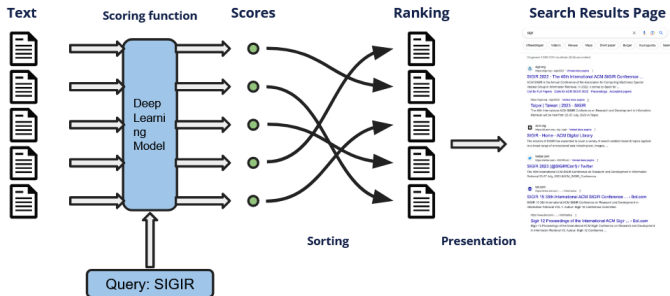
# Pemeringkatan Teks dengan *Machine Learning*

1. Kekurangan penggunaan *machine learning* untuk pemeringkatan adalah jumlah fitur tambahan yang dibutuhkan cukup banyak untuk mengimbangi kekurangan dari BM25, dan fitur tersebut biasanya dibuat secara manual (Lin, Nogueira, & Yates, 2020).
2. Batasan yang dialami model *machine learning* pada era *learning to rank*, diatasi dengan menggunakan *deep learning*.



# Alur Pemeringkatan Teks dengan *Deep Learning*

## Text Ranking With Deep Learning





# Rumusan Masalah

1. Bagaimana pengaplikasian model BERT untuk pemeringkatan teks berbahasa Indonesia?
2. Bagaimana kinerja model BERT pada setiap *dataset* yang digunakan bila dibandingkan dengan model *baseline* BM25?



# Tujuan Penelitian

1. Membangun dan melatih kembali (*fine tuning*) model BERT untuk pemeringkatan teks berbahasa Indonesia.
2. Membandingkan kinerja model BERT pada setiap *dataset* yang digunakan bila dibandingkan dengan model *baseline* BM25.



# Batasan Masalah

1. *Dataset* yang digunakan untuk melatih kembali (*fine tuning*) model BERT adalah *dataset* Mmarco *train set* bahasa Indonesia (Bonifacio, Campiotti, de Alencar Lotufo, & Nogueira, 2021).
2. *Dataset* yang digunakan untuk mengukur performa model adalah *dataset* Mmarco *dev set* bahasa Indonesia (Bonifacio et al., 2021) untuk *in-domain test* serta MrTyDi *dev set* bahasa Indonesia (Zhang, Ma, Shi, & Lin, 2021), dan Miracl *dev set* bahasa Indonesia (Zhang et al., 2023) untuk *out-of-domain test*.
3. Kinerja model diamati dengan metrik *reciprocal rank* (RR), *recall* (R), dan *normalized discounted cumulative gain* (NDCG).



# Table of Contents

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT





# Pemeringkatan Teks 1

## Tugas Pemeringkatan Teks

Diberikan kueri  $q$  dan himpunan teks terbatas  $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ , keluaran yang diinginkan dari permasalahan ini adalah barisan teks  $D_k = (d_{i_1}, d_{i_2}, \dots, d_{i_k})$  yang merupakan  $k$  teks yang paling relevan dengan kueri  $q$ .

*Dataset* Uji pada masalah pemeringkatan teks terdiri dari tiga *file*, yaitu *file* kueri, *file* korpus dan *file judgements*.



## Pemeringkatan Teks 2

Table: *File* korpus

<b>_id</b>	<b>title</b>	<b>text</b>
1342516#1	Colobothea biguttata	Larva kumbang ini biasanya mengebor ke dalam kayu dan dapat menyebabkan kerusakan ...
1342517#0	Ichthyodes rufipes	Ichthyodes rufipes adalah spesies kumbang tanduk panjang yang berasal dari famili Cerambycidae. Spesies ini ...



## Pemeringkatan Teks 3

Table: *File* kueri

<b>_id</b>	<b>text</b>
3	Dimana James Hepburn meninggal?
4	Dimana Jamie Richard Vardy lahir?
11	berapakah luas pulau Flores?
17	Siapa yang menulis Candy Candy?
19	Apakah karya tulis Irma Hardisurya yang pertama?



# Pemeringkatan Teks 4

Table: *File judgements*

query-id	corpus-id	score
3	115796#6	1
3	77689#48	1
4	1852373#0	1



# Table of Contents

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT



## Recall dan Presisi

$$\text{recall}(q, D_k)@k = \frac{\sum_{d \in D_k} \text{rel}(q, d)}{\sum_{d \in \mathcal{D}} \text{rel}(q, d)} \in [0, 1],$$

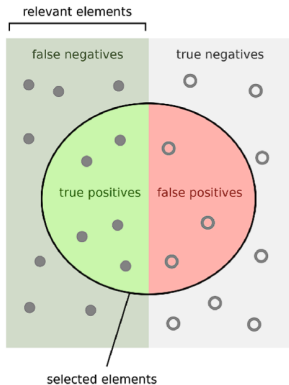
$$\text{precision}(q, D_k)@k = \frac{\sum_{d \in D_k} \text{rel}(q, d)}{|D_k|} \in [0, 1],$$

$$\text{rel}(q, d) = \begin{cases} 1 & \text{jika } r > 1 \\ 0 & \text{jika } r = 0 \end{cases}.$$

- ▶  $q$ : kueri,
- ▶  $D_k$ : barisan  $k$  teks yang dipilih oleh sistem,
- ▶  $r$ : nilai relevansi antara kueri  $q$  dengan teks  $d$  dari *file judgements*.



# Recall dan Presisi



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Figure: Ilustrasi *recall* dan presisi.



## Reciprocal Rank

Metrik lainnya yang sering digunakan untuk mengukur performa sistem pemeringkatan adalah *reciprocal rank* (RR). Metrik RR menitikberatkan pada peringkat dari teks relevan pertama dengan kueri  $q$ .

$$RR(q, D_k)@k = \begin{cases} \frac{1}{\text{FirstRank}(q, D_k)} & \text{jika } \exists d \in D_k \text{ dengan } \text{rel}(q, d) = 1 \\ 0 & \text{jika } \forall d \in D_k, \text{rel}(q, d) = 0 \end{cases},$$

- ▶  $q$ : kueri,
- ▶  $D_k$ : barisan  $k$  teks yang dipilih oleh sistem,
- ▶  $r$ : nilai relevansi antara kueri  $q$  dengan teks  $d$  dari *file judgements*.
- ▶  $\text{FirstRank}(q, D_k)$ :  
posisi teks relevan pertama  $d \in D_k$  dengan  $\text{rel}(q, d) = 1$ .





# Reciprocal Rank

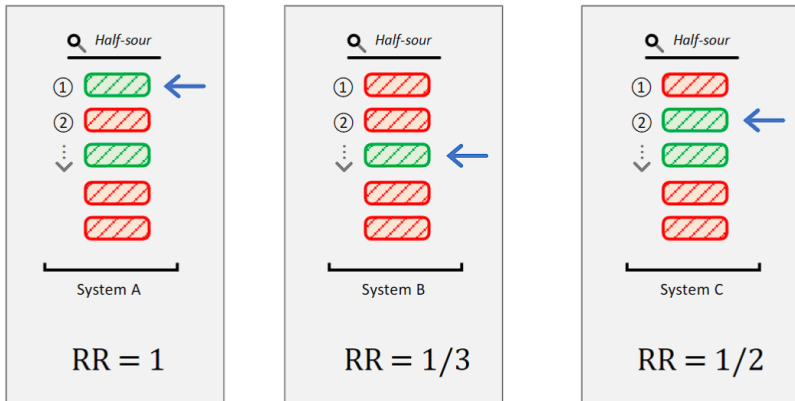


Figure: Ilustrasi *reciprocal rank*.



## Normalized Discounted Cumulative Gain

*Normalized Discounted Cumulative Gain* (NDCG) adalah metrik yang umumnya digunakan untuk mengukur kualitas dari pencarian situs web. Tidak seperti metrik yang telah disebutkan sebelumnya, nDCG dirancang untuk suatu  $r$  yang tak biner.

$$\text{nDCG}(q, D_k)@k = \frac{\text{DCG}(q, D_k)@k}{\text{DCG}(q, D_k^{\text{ideal}})@k} \in [0, 1],$$

$$\text{DCG}(q, D_k)@k = \sum_{d \in D_k} \frac{2^{\text{rel}(q, d)} - 1}{\log_2(\text{rank}(d, D_k) + 1)},$$

$\text{rank}(d, D_k)$  = Posisi  $d$  dalam  $D_k$ ,

$\text{rel}(q, d) = r$ .



## Normalized Discounted Cumulative Gain

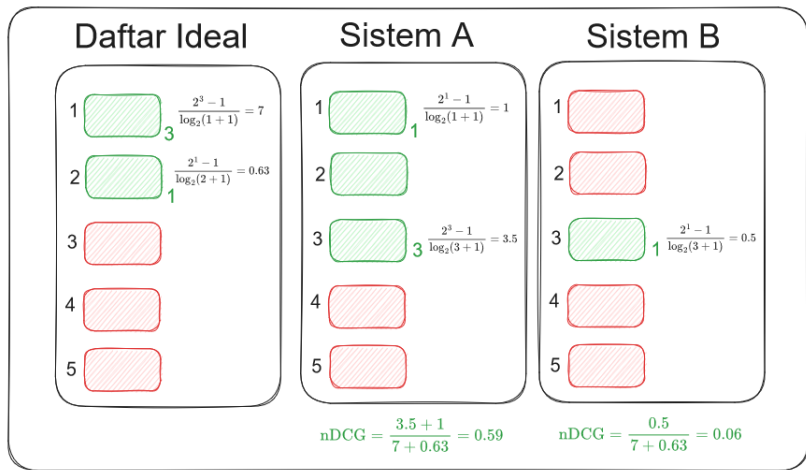


Figure: Ilustrasi *normalized discounted cumulative gain*.



# Table of Contents

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT



# Pemeringkatan Teks Dengan Statistik

1. Untuk mengambil  $k$  teks dari kumpulan  $\mathcal{D}$ , kita menggunakan fungsi skor  $\text{score}(q, d, \mathcal{D})$  untuk mengukur relevansi antara kueri  $q$  dan teks  $d$ . Dengan mencari skor antara  $q$  dan semua teks pada  $\mathcal{D}$ , kita dapat memilih barisan teks  $D_k = (d_{i_1}, d_{i_2}, \dots, d_{i_k})$  dengan  $k$  teks memiliki skor tertinggi.
2. Salah satu fungsi skor mudah dan sering digunakan adalah TF-IDF dan BM25. Fungsi skor ini menghitung skor antara kueri  $q$  dan teks  $d$  dengan informasi dari kata yang ada pada  $q$  dan  $d$ .



# TF-IDF

- ▶ *term frequency*:  $tf(t, d) = \frac{\text{Count}(t, d)}{|d|}$ ,
- ▶ *document frequency*:  
 $df(t, \mathcal{D}) = \text{jumlah teks pada } \mathcal{D} \text{ yang mengandung kata } t$ .
- ▶ *inverse document frequency*:  
$$idf(t, \mathcal{D}) = \begin{cases} \log_2 \left( \frac{|\mathcal{D}|}{df(t, \mathcal{D})} \right) & \text{jika } df(t, \mathcal{D}) > 0 \\ 0 & \text{jika } df(t, \mathcal{D}) = 0 \end{cases}$$
- ▶  $TF\text{-}IDF(t, d, \mathcal{D}) = tf(t, d) \times idf(t, \mathcal{D})$ .

	doc <sub>1</sub>	doc <sub>2</sub>	doc <sub>3</sub>	doc <sub>4</sub>
A	10	10	10	10
B	10	10	10	0
C	10	10	0	0
D	0	0	0	1

⇒

	IDF
A	0.00
B	0.29
C	0.69
D	1.39

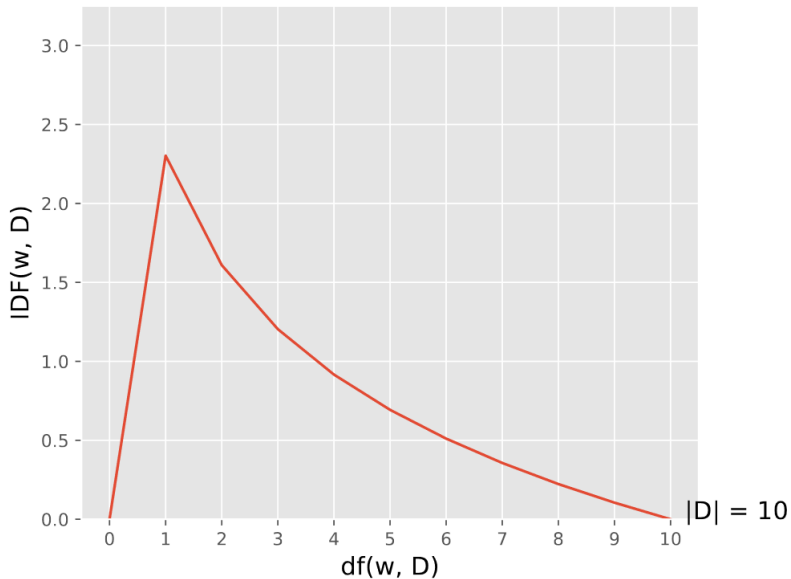


	TF			
	doc <sub>1</sub>	doc <sub>2</sub>	doc <sub>3</sub>	doc <sub>4</sub>
A	0.33	0.33	0.50	0.91
B	0.33	0.33	0.50	0.00
C	0.33	0.33	0.00	0.00
D	0.00	0.00	0.00	0.09

	TF-IDF			
	doc <sub>1</sub>	doc <sub>2</sub>	doc <sub>3</sub>	doc <sub>4</sub>
A	0.00	0.00	0.00	0.00
B	0.10	0.10	0.14	0.00
C	0.23	0.23	0.00	0.00
D	0.00	0.00	0.00	0.13



# Nilai IDF



# Score

$$\text{score}(q, d, \mathcal{D}) = \sum_{t \in T_q \cap T_d} \text{TF-IDF}(t, d, \mathcal{D})$$

$T_q = \{t_1, t_2, \dots, t_{L_1}\}$  = kumpulan kata pada  $q$ ,

$T_d = \{t_1, t_2, \dots, t_{L_2}\}$  = kumpulan kata pada  $d$ .





# BM25

## *Smoothed* IDF

$$\text{idf}_{\text{BM25}}(t, \mathcal{D}) = \log \left( 1 + \frac{|\mathcal{D}| - \text{df}(t, \mathcal{D}) + 0.5}{\text{df}(t, \mathcal{D}) + 0.5} \right)$$

## Score BM25 Pengganti tf

$$\text{score}_{\text{BM25}}(t, d) = \frac{\text{tf}(t, d) \times (k_1 + 1)}{\text{tf}(t, d) + k_1 \times (1 - b + b \times \frac{|d|}{\text{avgdl}})}$$

## BM25

$$\text{BM25}(t, d, \mathcal{D}) = \text{idf}_{\text{BM25}}(t, \mathcal{D}) \times \text{score}_{\text{BM25}}(q, d, \mathcal{D})$$

Robertson, Walker, Jones, Hancock-Beaulieu, dan Gatford (1994)



# Table of Contents

Pendahuluan

Pemeringkatan Teks

Metrik Evaluasi

Pemeringkatan Teks Dengan Statistik

Mekanisme *Attention*, *Transformer* dan BERT



# Mekanisme *Attention*



# *Soft Attention*



## Attention Parametrik



# Transformer





# Daftar Pustaka I

- Bonifacio, L. H., Campiotti, I., de Alencar Lotufo, R., & Nogueira, R. F. (2021). mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, *abs/2108.13897*. Diakses dari <https://arxiv.org/abs/2108.13897>
- Lin, J., Nogueira, R. F., & Yates, A. (2020). Pretrained transformers for text ranking: BERT and beyond. *CoRR*, *abs/2010.06467*. Diakses dari <https://arxiv.org/abs/2010.06467>
- Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at trec-3. In *Text retrieval conference*. Diakses dari <https://api.semanticscholar.org/CorpusID:3946054>
- Zhang, X., Ma, X., Shi, P., & Lin, J. (2021). Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv:2108.08787*.





## Daftar Pustaka II

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., ... Lin, J. (2023, 09). MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11, 1114-1131. Diakses dari [https://doi.org/10.1162/tac1\\_a\\_00595](https://doi.org/10.1162/tac1_a_00595) doi: 10.1162/tac1\_a\_00595

