

#### **UNIVERSITAS INDONESIA**

#### PEMERINGKATAN TEKS BAHASA INDONESIA DENGAN BERT

#### **SKRIPSI**

# CARLES OCTAVIANUS 2006568613

# FAKULTAS FAKULTAS MATEMATIKA DAN ILMU PENGATAHUAN ALAM PROGRAM STUDI MATEMATIKA DEPOK DESEMBER 2023



#### **UNIVERSITAS INDONESIA**

#### PEMERINGKATAN TEKS BAHASA INDONESIA DENGAN BERT

#### **SKRIPSI**

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains

CARLES OCTAVIANUS 2006568613

FAKULTAS FAKULTAS MATEMATIKA DAN ILMU PENGATAHUAN ALAM
PROGRAM STUDI MATEMATIKA
DEPOK
DESEMBER 2023

#### HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : Carles Octavianus

NPM : 2006568613

Tanda Tangan :

Tanggal : 2 Desember 2023

#### HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Carles Octavianus

NPM : 2006568613 Program Studi : Matematika

Judul Skripsi : Pemeringkatan Teks Bahasa Indonesia Dengan BERT

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana pada Program Studi Matematika, Fakultas Fakultas Matematika dan Ilmu Pengatahuan Alam, Universitas Indonesia.

#### **DEWAN PENGUJI**

Pembimbing 1	:	Sarini Abdullah S.Si., M.Stats., Ph.D.	(	)
Penguji 1	:	Penguji Pertama Anda	(	)
Penguji 2	:	Penguji Kedua Anda	(	)

Ditetapkan di : Depok

Tanggal : 2 Desember 2023

#### **KATA PENGANTAR**

Template ini disediakan untuk orang-orang yang berencana menggunakan LAT<sub>E</sub>X untuk membuat dokumen tugas akhir.

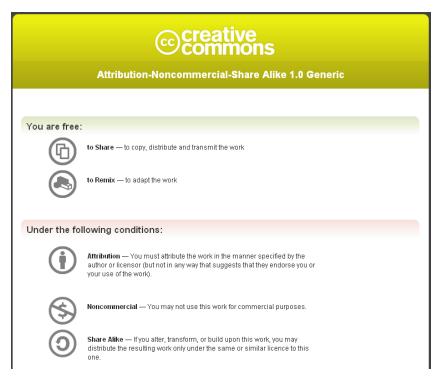
#### @todo

Silakan ganti pesan ini dengan pendahuluan kata pengantar Anda.

#### Ucapan Terima Kasih:

- 1. Pembimbing.
- 2. Dosen.
- 3. Instansi.
- 4. Orang tua.
- 5. Sahabat.
- 6. Teman.

Penulis menyadari bahwa laporan Skripsi ini masih jauh dari sempurna. Oleh karena itu, apabila terdapat kesalahan atau kekurangan dalam laporan ini, Penulis memohon agar kritik dan saran bisa disampaikan langsung melalui *e-mail* emailanda@mail.id.



Creative Common License 1.0 Generic

Terkait template ini, gambar lisensi di atas diambil dari http://creativecommons.org/licenses/by-nc-sa/1.0/deed.en\_CA. Jika ingin mengentahui lebih lengkap mengenai *Creative Common License 1.0 Generic*, silahkan buka http://creativecommons.org/licenses/by-nc-sa/1.0/legalcode. Seluruh dokumen yang dibuat dengan menggunakan template ini sepenuhnya menjadi hak milik pembuat dokumen dan bebas didistribusikan sesuai dengan keperluan masing-masing. Lisensi hanya berlaku jika ada orang yang membuat template baru dengan menggunakan template ini sebagai dasarnya.

Penyusun template ingin berterima kasih kepada Andreas Febrian, Lia Sadita, Fahrurrozi Rahman, Andre Tampubolon, dan Erik Dominikus atas kontribusinya dalam template yang menjadi pendahulu template ini. Penyusun template juga ingin mengucapkan terima kasih kepada Azhar Kurnia atas kontribusinya dalam template yang menjadi pendahulu template ini.

Semoga template ini dapat membantu orang-orang yang ingin mencoba menggunakan IATEX. Semoga template ini juga tidak berhenti disini dengan ada kontribusi dari para penggunanya. Jika Anda memiliki perubahan yang dirasa penting untuk disertakan dalam template, silakan lakukan *fork* repositori Git template ini di https://gitlab.com/ichlaffterlalu/latex-skripsi-ui-2017, lalu lakukan *merge request* 

perubahan Anda terhadap *branch* master. Kami berharap agar *template* ini dapat terus diperbarui mengikuti perubahan ketentuan dari pihak Rektorat Universitas Indonesia, dan hal itu tidak mungkin terjadi tanpa kontribusi dari teman-teman sekalian.

Depok, 2 Desember 2023

Carles Octavianus

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS

AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama : Carles Octavianus

NPM : 2006568613

Program Studi : Matematika

**Jenis Karya** : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif** (*Non-exclusive Royalty Free Right*) atas

karya ilmiah saya yang berjudul:

Pemeringkatan Teks Bahasa Indonesia Dengan BERT

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif

ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola

dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir saya

selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik

Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok

Pada tanggal : 2 Desember 2023

Yang menyatakan

(Carles Octavianus)

#### **ABSTRAK**

Nama : Carles Octavianus

Program Studi : Matematika

Judul : Pemeringkatan Teks Bahasa Indonesia Dengan BERT

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

Isi abstrak.

Kata kunci:

Keyword satu, kata kunci dua

#### **ABSTRACT**

Name : Carles Octavianus

Study Program : Mathematics

Title : Text Ranking in Indonesian Using BERT Counselor : Sarini Abdullah S.Si., M.Stats., Ph.D.

Abstract content.

Key words:

Keyword one, keyword two

# DAFTAR ISI

H	ALAN	MAN JUDUL	i
LE	EMBA	AR PENGESAHAN	ij
K	ATA P	PENGANTAR ii	ii
LE	EMBA	AR PERSETUJUAN PUBLIKASI ILMIAH	ر ز
ΑI	BSTR	AK	ii
DA	AFTA	R ISI	X
DA	<b>A</b> FTA	R GAMBAR	j
DA	AFTA	R TABEL	ij
DA	AFTA	R KODE PROGRAM xii	ij
DA	<b>A</b> FTA	R LAMPIRAN	V
1	PEN	DAHULUAN	1
2	LAN	NDASAN TEORI	2
3		IRECTIONAL ENCODER REPRESENTATION FROM TRANS-	_
			3
	3.1	v	3
		v ·	<i>3</i>
			3
			3
			3
	3.2	D'I' I' IE I D II (DEDT)	_
	3.2		<i>3</i>
		1	3
			3
			3
	3.3		3
	3.4	***	3
4			4
	4. I	Spesifikasi Mesin dan Perangkat Lunak	4

	4.2	Tahapa	n Simulas	si													4
	4.3	Dataset	t Latih daı	ı Uji													5
		4.3.1	Dataset I	Latih													5
			4.3.1.1	Mmarce	o Indon	esia '	Tra	in	Set	t.							5
		4.3.2	Dataset U	J <b>ji</b> .													5
			4.3.2.1	Mmarce	o Indon	esia	DE	V S	Set								5
			4.3.2.2	Mrtydi	Indones	sia T	ES.	ΓS	et								5
			4.3.2.3	Miracl	Indones	sia Tl	EST	$\Gamma$ S	et								5
	4.4	Metriks	s Evaluasi														5
	4.5	Fine Tu	ining BEF	RT													5
		4.5.1	IndoBER	$T_{CAT}$													5
		4.5.2	IndoBER	$T_{\mathrm{DOT}}$													5
		4.5.3	IndoBER														5
		4.5.4	IndoBER		-												5
		4.5.5	IndoBER														5
	4.6	Hasil F	ine Tunin														5
		4.6.1	Evaluasi	BM25													5
		4.6.2	Evaluasi	IndoBEI	RT <sub>MEAN</sub>	1											6
		4.6.3	Evaluasi	IndoBEI	$RT_{CAT}$												6
		4.6.4	Evaluasi														6
		4.6.5	Evaluasi	IndoBEI	RT <sub>DOTh</sub>	ardneg	s ·										6
		4.6.6	Evaluasi														7
		4.6.7	Evaluasi														7
		4.6.8	Perbandi	ngan Ha	sil Eval	uasi		•		•							7
5	PEN	UTUP													• •		9
	5.1		oulan														9
	5.2	-															9
D.	A E/TA	D DEFI	DENEL														10

## DAFTAR GAMBAR

# DAFTAR TABEL

Caption	5
Caption	6
Caption	7
Caption	7
Caption	7
Caption	8
	Caption

# DAFTAR KODE PROGRAM

# DAFTAR LAMPIRAN

Lampiran 1.	CHANGELOG	12
Lampiran 2.	Judul Lampiran 2	14

# BAB 1 PENDAHULUAN

@todo			
wew			

# BAB 2 LANDASAN TEORI

@todo			

#### **BAB 3**

# BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER (BERT) UNTUK PEMERINGKATAN TEKS

# @todo jabarin sih isinya mau gmna

- 3.1 Transformer
- 3.1.1 Attention
- 3.1.1.1 Self-Attention
- 3.1.1.2 Multi-Head Attention
- 3.1.2 Positional Encoding
- 3.1.3 Feed Forward Network
- 3.1.4 Arsitektur *Transformer* Enkoder
- 3.2 Bidirectional Encoder Representation from Transformer (BERT)
- 3.2.1 model pralatih BERT
- 3.2.1.1 Tugas Masked Language Model (MLM)
- **3.2.1.2** Tugas Next Sentence Prediction (NSP)
- 3.2.2 IndoBERT
- 3.3 Arsitektur BERT<sub>CAT</sub> Untuk Pemeringkatan Teks
- 3.4 Arsitektur BERT<sub>DOT</sub> Untuk Pemeringkatan Teks

#### BAB 4

#### HASIL SIMULASI DAN PEMBAHASAN

Bab ini membahas mengenai proses fine tuning model Bidirectional Encoder Representations from Transformers (BERT) untuk mendapatkan model yang dapat digunakan untuk masalah pemeringkatan teks. Subbab4.1 menjelaskan mengenai spesifikasi perangkat keras dan perangkat lunak yang digunakan dalam penelitian. Selanjutnya, Subbab 4.2 menjelaskan mengenai tahapan simulasi yang dilakukan dalam penelitian. Dataset latih (train) dan uji (validation) dijelaskan pada Subbab 4.3. Subbab 4.5 menjelaskan lebih detail mengenai arsitektur model BERT,fungsi loss, serta konfigurasi hyperparameter yang digunakan dalam proses fine tuning model BERT. Subbab 4.4 menjelaskan kembali mengenai metriks evaluasi yang digunakan pada setiap dataset uji yang digunakan. Terakhir, Subbab 4.6 menjelaskan mengenai hasil fine tuning model BERT dan evaluasi dari modelmodel yang dihasilkan.

#### 4.1 Spesifikasi Mesin dan Perangkat Lunak

#### @todo

banyak sih: 'D, tambahin tabel isi qid, pid, label buat mmarco train tambahin tabel isi qid, pid, label buat miracl test, tunjukkin ini lebih dense dari mrtydi dan mmarco dev/ mrtydi test

Proses fine tuning model BERT untuk pemeringkatan teks dilakukan menggunakan mesin dan perangkat lunak yang tertera pada berikut.

#### 4.2 Tahapan Simulasi

menunjukkan tahapan simulasi yang dilakukan dalam penelitian ini.

#### 4.3 Dataset Latih dan Uji

#### 4.3.1 Dataset Latih

#### 4.3.1.1 Mmarco Indonesia Train Set

#### 4.3.2 Dataset Uji

- 4.3.2.1 Mmarco Indonesia DEV Set
- 4.3.2.2 Mrtydi Indonesia TEST Set
- 4.3.2.3 Miracl Indonesia TEST Set

#### 4.4 Metriks Evaluasi

#### 4.5 Fine Tuning BERT

#### 4.5.1 IndoBERT<sub>CAT</sub>

#### 4.5.2 IndoBERT<sub>DOT</sub>

# 4.5.3 IndoBERT<sub>DOThardnegs</sub>

# 4.5.4 IndoBERT<sub>DOTMargin</sub>

#### 4.5.5 IndoBERT<sub>KD</sub>

#### 4.6 Hasil Fine Tuning dan Evaluasi

#### **4.6.1** Evaluasi BM25

Tabel 4.1: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971

Tabel 4.2: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.1 <u>1</u> 4	.642	.279	.858	.391	.971
IndoBERT <sub>MEAN</sub>	.000	.000	.000	.000	.000	.000

# $4.6.2 \quad Evaluasi \ IndoBERT_{MEAN}$

# 4.6.3 Evaluasi IndoBERT<sub>CAT</sub>

Tabel 4.3: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>CAT</sub>	.181	.642	.447	.858	.455	.971

## 4.6.4 Evaluasi IndoBERT<sub>DOT</sub>

Tabel 4.4: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>DOT</sub>	.192	.847	.378	.936	.355	.920

# 4.6.5 Evaluasi IndoBERT<sub>DOThardnegs</sub>

Tabel 4.5: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>DOThardnegs</sub>	.232	.847	.471	.921	.397	.898

# 4.6.6 Evaluasi IndoBERT<sub>DOTMargin</sub>

Tabel 4.6: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>DOTMargin</sub>	.207	.799	.446	.929	.387	.899

# 4.6.7 Evaluasi IndoBERT<sub>KD</sub>

Tabel 4.7: Caption

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>KD</sub>	-	.803	.300	.761	-	-

# 4.6.8 Perbandingan Hasil Evaluasi

Tabel 4.8: Caption

Model	Mmarc	o Dev	MrTyD	MrTyDi Test Miracl Dev		Dev
	MRR@10	R@1000	MRR@10	R@1000	NCDG@10	R@1K
BM25 (Elastic Search)	.114	.642	.279	.858	.391	.971
IndoBERT <sub>MEAN</sub>	.000	.000	.000	.000	.000	.000
IndoBERT <sub>CAT</sub>	.181	.642	.447	.858	.455	.971
IndoBERT <sub>DOT</sub>	.192	.847	.378	.936	.355	.920
IndoBERT <sub>DOTdnegs</sub>	.232	.847	.471	.921	.397	.898
IndoBERT <sub>DOTMargin</sub>	.207	.799	.446	.929	.387	.899
IndoBERT <sub>KD</sub>	-	.803	.300	.761	-	-

Tabel 4.9: Caption

Model	Latensi (ms)	Memori(MB)
BM25 (Elastic Search)	6.55	800
IndoBERT <sub>DOT</sub>	9.9	3072
IndoBERT <sub>CAT</sub>	242	800

#### **BAB 5**

#### **PENUTUP**

Pada bab ini, Penulis akan memaparkan kesimpulan penelitian dan saran untuk penelitian berikutnya.

#### 5.1 Kesimpulan

Berikut ini adalah kesimpulan terkait pekerjaan yang dilakukan dalam penelitian ini:

#### 1. Poin pertama

Penjelasan poin pertama.

#### 2. Poin kedua

Penjelasan poin kedua.

Tulis kalimat penutup di sini.

#### 5.2 Saran

Berdasarkan hasil penelitian ini, berikut ini adalah saran untuk pengembangan penelitian berikutnya:

- 1. Saran 1.
- 2. Saran 2.

#### **DAFTAR REFERENSI**



#### **LAMPIRAN 1: CHANGELOG**

#### @todo

Silakan hapus lampiran ini ketika Anda mulai menggunakan template.

*Template* versi terbaru bisa didapatkan di https://gitlab.com/ichlaffterlalu/latex-skripsi-ui-2017. Daftar perubahan pada *template* hingga versi ini:

- versi 1.0.3 (3 Desember 2010):
  - Template Skripsi/Tesis sesuai ketentuan formatting tahun 2008.
  - Bisa diakses di https://github.com/edom/uistyle.
- versi 2.0.0 (29 Januari 2020):
  - Template Skripsi/Tesis sesuai ketentuan formatting tahun 2017.
  - Menggunakan BibTeX untuk sitasi, dengan format default sitasi IEEE.
  - Template kini bisa ditambahkan kode sumber dengan code highlighting untuk bahasa pemrograman populer seperti Java atau Python.
- versi 2.0.1 (8 Mei 2020):
  - Menambahkan dan menyesuaikan tutorial dari versi 1.0.3, beserta cara kontribusi ke template.
- versi 2.0.2 (14 September 2020):
  - Versi ini merupakan hasil feedback dari peserta skripsi di lab Reliable Software Engineering (RSE) Fasilkom UI, semester genap 2019/2020.
  - BibTeX kini menggunakan format sitasi APA secara default.
  - Penambahan tutorial untuk longtable, agar tabel bisa lebih dari 1 halaman dan header muncul di setiap halaman.
  - Menambahkan tutorial terkait penggunaan BibTeX dan konfigurasi header/footer untuk pencetakan bolak-balik.

- Label "Universitas Indonesia" kini berhasil muncul di halaman pertama tiap
   bab dan di bagian abstrak daftar kode program.
- Hyphenation kini menggunakan babel Bahasa Indonesia. Aktivasi dilakukan di hype-indonesia.tex.
- Minor adjustment untuk konsistensi *license* dari template.
- versi 2.0.3 (15 September 2020):
  - Menambahkan kemampuan orientasi *landscape* beserta tutorialnya.
  - \captionsource telah diperbaiki agar bisa dipakai untuk longtable.
  - Daftar lampiran kini telah tersedia, lampiran sudah tidak masuk daftar isi lagi.
  - Nomor halaman pada lampiran dilanjutkan dari halaman terakhir konten (daftar referensi).
  - Kini sudah bisa menambahkan daftar isi baru untuk jenis objek tertentu (custom), seperti: "Daftar Aturan Transformasi". Sudah termasuk mekanisme captioning dan tutorialnya.
  - Perbaikan minor pada tutorial.
- versi 2.1.0 (8 September 2021):
  - Versi ini merupakan hasil feedback dari peserta skripsi dan tesis di lab Reliable
     Software Engineering (RSE) Fasilkom UI, semester genap 2020/2021.
  - Minor edit: "Lembar Pengesahan", dsb. di daftar isi menjadi all caps.
  - Experimental multi-language support (Chinese, Japanese, Korean).
  - Support untuk justifikasi dan word-wrapping pada tabel.
  - Penggunaan suffix "(sambungan)" untuk tabel lintas halaman. Tambahan support suffix untuk \captionsource.
- versi 2.1.1 (7 Februari 2022):
  - Update struktur mengikuti fork template versi 1.0.3 di https://github
     .com/rkkautsar/edom/ui-thesis-template.
  - Support untuk simbol matematis amsfonts.

- Kontribusi komunitas terkait improvement GitLab CI, atribusi, dan format sitasi APA bahasa Indonesia.
- Perbaikan tutorial berdasarkan perubahan terbaru pada versi 2.1.0 dan 2.1.1.
- versi 2.1.2 (13 Agustus 2022):
  - Modifikasi penamaan beberapa berkas.
  - Perbaikan beberapa halaman depan (halaman persetujuan, halaman orisinalitas, dsb.).
  - Support untuk lembar pengesahan yang berbeda dengan format standar, seperti Laporan Kerja Praktik dan Disertasi.
  - Kontribusi komunitas terkait kesesuaian dengan format Tugas Akhir UI, kelengkapan dokumen, perbaikan format sitasi, dan *quality-of-life*.
  - Perbaikan tutorial.
- versi 2.1.3 (22 Februari 2023):
  - Dukungan untuk format Tugas Akhir Kelompok di Fasilkom UI.
  - Dukungan untuk format laporan Kampus Merdeka Mandiri di Fasilkom UI.
  - Minor bugfix: Perbaikan kapitalisasi variabel.
  - Quality-of-Life: Pengaturan kembali config/settings.tex.
  - Tutorial untuk beberapa use case.

#### LAMPIRAN 2: JUDUL LAMPIRAN 2

Lampiran hadir untuk menampung hal-hal yang dapat menunjang pemahaman terkait tugas akhir, namun akan mengganggu *flow* bacaan sekiranya dimasukkan ke dalam bacaan. Lampiran bisa saja berisi data-data tambahan, analisis tambahan, penjelasan istilah, tahapan-tahapan antara yang bukan menjadi fokus utama, atau pranala menuju halaman luar yang penting.

#### Subbab dari Lampiran 2

#### @todo

Isi subbab ini sesuai keperluan Anda. Anda bisa membuat lebih dari satu judul lampiran, dan tentunya lebih dari satu subbab.