



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 Matematika dan Ilmu Pengetahuan 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 L^AT_EX 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`.

Terkait template ini, gambar lisensi di atas diambil dari <http://creativecommons.org/licenses/by-nc-sa/1.0/deed.en-CA>. Jika ingin mengetahui 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, Fahrur-rozi 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 menggu-

nakan L^AT_EX. 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

demikian 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

berserta 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

HALAMAN JUDUL	i
LEMBAR PENGESAHAN	ii
KATA PENGANTAR	iii
LEMBAR PERSETUJUAN PUBLIKASI ILMIAH	v
ABSTRAK	vi
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xiii
DAFTAR KODE PROGRAM	xv
DAFTAR LAMPIRAN	xvi
1 PENDAHULUAN	1
2 LANDASAN TEORI	2
2.1 Masalah Pemeringkatan Teks	2
2.1.1 Bentuk Umum <i>Dataset</i>	2
2.1.2 Metrik Evaluasi dalam Pemeringkatan Teks	4
2.1.2.1 <i>Recall</i> dan Presisi	4
2.1.2.2 <i>Reciprocal Rank</i>	6
2.1.2.3 <i>Normalized Discounted Cumulative Gain</i> (nDCG)	7
2.2 Pemeringkatan Teks dengan Statistik	8
2.2.1 <i>Term Frequency - Inverse Document Frequency</i> (TF-IDF)	8
2.2.2 <i>Best Match 25</i> (BM25)	11
2.3 <i>Deep Learning</i>	14
2.3.1 <i>Multilayer Perceptron</i> (MLP)	15
2.3.2 Fungsi Aktivasi	16
2.3.3 Fungsi <i>Loss</i>	17
2.3.4 <i>Optimasi</i> Parameter	19
2.4 Pembelajaran Representasi	23
2.4.1 Fungsi <i>Loss</i> pada Pembelajaran Representasi	23
3 BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANS-	
FORMER (BERT) UNTUK PEMERINGKATAN TEKS	25
3.1 Mekanisme <i>Attention</i>	25

3.1.1	<i>Attention</i> Parametrik	27
3.2	Transformer	28
3.2.1	<i>Token Embedding (Input Embedding)</i>	30
3.2.2	<i>Scaled Dot-Product Attention</i>	31
3.2.3	<i>Self-Attention</i>	33
3.2.4	<i>Multi-Head Self-Attention</i>	35
3.2.5	<i>Positional Encoding</i>	36
3.2.6	<i>Position-wise Feed-Forward Network</i>	37
3.2.7	Koneksi Residu dan <i>Layer Normalization</i>	38
3.2.8	Transformer Encoder	40
3.3	<i>Bidirectional Encoder Representations from Transformers (BERT)</i>	41
3.3.1	Representasi Input	41
3.3.2	<i>pre-training BERT</i>	41
3.3.2.1	<i>Masked Language Model (MLM)</i>	42
3.3.2.2	<i>Next Sentence Prediction</i>	43
3.3.3	BERT untuk Bahasa Indonesia (IndoBERT)	43
3.3.4	Penggunaan BERT untuk Peningkatan Teks	44
3.3.4.1	BERT _{CAT}	44
3.3.4.2	BERT _{DOT}	46
4	HASIL SIMULASI DAN PEMBAHASAN	48
4.1	Spesifikasi Mesin dan Perangkat Lunak	48
4.2	Tahapan Simulasi	48
4.3	Data	50
4.4	<i>Fine Tuning Model BERT</i>	51
4.4.1	IndoBERT _{CAT}	51
4.4.2	IndoBERT _{DOT}	53
4.4.3	IndoBERT _{DOTHardnegs}	54
4.4.4	IndoBERT _{DOTKD}	55
4.5	Evaluasi Model	57
4.5.1	Evaluasi IndoBERT _{CAT}	57
4.5.2	Evaluasi IndoBERT _{DOT}	58
4.5.3	Evaluasi IndoBERT _{DOTHardnegs}	59
4.5.4	Evaluasi IndoBERT _{KD}	59
4.6	Diskusi Hasil	60
5	PENUTUP	64
5.1	Kesimpulan	64
5.2	Saran	64
	DAFTAR REFERENSI	65

DAFTAR GAMBAR

Gambar 2.1.	Ilustrasi <i>recall</i> dan presisi. Nilai <i>recall</i> dihitung sebagai rasio teks relevan yang diambil oleh sistem terhadap seluruh teks yang relevan dengan kueri q . Sedangkan nilai presisi dihitung sebagai rasio teks relevan yang diambil oleh sistem terhadap seluruh teks yang diambil oleh sistem.	4
Gambar 2.2.	Ilustrasi <i>reciprocal rank</i> (RR). Nilai RR pada sistem A, B, dan C berturut-turut adalah 1, 0.33, dan 0.5 karena posisi dari teks yang relevan pertama adalah 1, 3, dan 2.	6
Gambar 2.3.	Ilustrasi perhitungan nDCG. Nilai nDCG dari sistem A, B, adalah rasio antara DCG dari sistem tersebut dengan DCG dari sistem ideal.	7
Gambar 2.4.	Grafik dari fungsi idf. Nilai idf menurun seiring dengan bertambahnya nilai $df(t, \mathcal{D})$	10
Gambar 2.5.	Ilustrasi perhitungan skor TF-IDF untuk suatu kumpulan teks.	11
Gambar 2.6.	Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai avgdl	12
Gambar 2.7.	Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai b	13
Gambar 2.8.	Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai k_1	13
Gambar 2.9.	Perbedaan antara idf_{BM25} dan idf.	14
Gambar 2.10.	Ilustrasi dari <i>Directed Acyclic Graph</i> (DAG) pada arsitektur <i>deep learning feed-forward neural network</i> (FFN).	14
Gambar 2.11.	Grafik dari fungsi aktivasi pada Tabel 2.4 dan turunannya.	17
Gambar 2.12.	Ilustrasi dari <i>Directed Acyclic Graph</i> (DAG) pada model <i>deep learning</i> ketika proses pelatihan dilakukan.	19
Gambar 2.13.	100 iterasi pertama dari pembaruan parameter $\theta = \{w_1, w_2\}$ dengan <i>learning rate</i> yang terlalu kecil dan terlalu besar.	20
Gambar 2.14.	100 iterasi pertama dari pembaruan parameter $\theta = \{w_1, w_2\}$ dengan <i>learning rate</i> yang baik.	21
Gambar 2.15.	Ilustrasi dari pembaruan parameter θ dengan <i>stochastic gradient descent</i> dengan momentum.	21
Gambar 2.16.	Ilustrasi dari pembaruan parameter θ dengan <i>stochastic gradient descent</i> dengan <i>adaptive learning rate</i>	22
Gambar 2.17.	Ilustrasi dari pembaruan parameter θ dengan Adam.	23
Gambar 2.18.	Ilustrasi fungsi objektif <i>N-pair loss</i> . Untuk pasangan teks yang relevan (a, b_1) , tujuannya adalah untuk meminimalkan jarak antara a dan b_1 sehingga jarak tersebut lebih kecil dibandingkan dengan jarak antara a dan b_i yang lain.	24
Gambar 3.1.	Ilustrasi dari mekanisme <i>soft attention</i>	26
Gambar 3.2.	Arsitektur <i>transformer</i> (Weng, 2018).	28

Gambar 3.3.	Ilustrasi dari representasi token. Gambar kiri menunjukkan representasi token dengan <i>one-hot encoding</i> , sedangkan gambar kanan menunjukkan representasi token dengan <i>token embedding</i>	30
Gambar 3.4.	Perbandingan RNN dan <i>self-attention</i> dalam menghasilkan representasi vektor kontekstual. Pada RNN, representasi vektor kontekstual setiap token bergantung pada perhitungan token sebelumnya. Pada <i>self-attention</i> , representasi vektor kontekstual setiap token dihitung secara independen dan paralel.	33
Gambar 3.5.	Ilustrasi <i>self-attention</i> dalam menghasilkan representasi vektor kontekstual dari barisan token. Representasi vektor dari token <i>it</i> akan bergantung terhadap barisan token <i>input</i>	34
Gambar 3.6.	Ilustrasi <i>multi-head self-attention</i> pada <i>transformer</i> . <i>Multi-head self-attention</i> menghitung <i>self-attention</i> sebanyak <i>h</i> kali pada subruang yang berbeda.	35
Gambar 3.7.	Ilustrasi dari <i>positional encoding</i> pada <i>transformer</i> . <i>Positional encoding</i> ditambahkan pada <i>token embedding</i> sebelum dijadikan <i>input</i> untuk <i>transformer</i>	37
Gambar 3.8.	Ilustrasi <i>position-wise feed-forward network</i> pada <i>transformer</i>	37
Gambar 3.9.	Ilustrasi koneksi residu.	39
Gambar 3.10.	Ilustrasi <i>transformer encoder</i> . (a) <i>Transformer encoder</i> , (b) <i>encoder blok</i>	41
Gambar 3.11.	Ilustrasi <i>Masked Language Modeling</i> (MLM) pada BERT. sebuah kata (token) secara acak di-hilangkan (<i>mask</i>) dan model diminta untuk menebak kata yang dihilangkan tersebut.	42
Gambar 3.12.	Ilustrasi <i>next sentence prediction</i> pada BERT. Model diminta untuk memprediksi apakah kalimat kedua adalah kalimat berikutnya dari kalimat pertama.	43
Gambar 3.13.	BERT _{CAT} mengambil kueri dan kandidat teks yang akan diberi skor sebagai <i>input</i> dan menggunakan BERT untuk klasifikasi relevansi. Penjumlahan elemen-wise dari token, <i>segment</i> , dan <i>positional embeddings</i> membentuk representasi vektor <i>input</i> . Setiap token input memiliki vektor kontekstual sebagai <i>output</i> model BERT. <i>Linear layer</i> menerima representasi akhir token [CLS] dan menghasilkan skor relevansi teks terkait dengan kueri (Lin, Nogueira, & Yates, 2020).	44
Gambar 3.14.	Arsitektur <i>retrieve and rerank</i> . <i>First-stage retrieval</i> dilakukan oleh BM25 dan <i>reranking</i> dilakukan oleh model <i>scoring</i> yang lebih kompleks seperti BERT _{CAT} (Hofstätter, Althammer, Sertkan, & Hanbury, 2021).	45
Gambar 3.15.	BERT _{DOT} memetakan kueri dan kandidat teks ke dalam ruang vektor yang sama dan menghitung skor relevansi dengan melakukan <i>dot product</i> antara vektor representasi kontekstual dari kueri dan teks (Hofstätter et al., 2021).	46
Gambar 3.16.	Arsitektur pemeringkatan dengan BERT _{DOT} . Vektor representasi dari setiap teks dapat diindeks terlebih dahulu dan disimpan dalam memori (Hofstätter et al., 2021).	47

Gambar 4.1.	Diagram Simulasi	49
Gambar 4.2.	Ilustrasi dari pelatihan model IndoBERT _{DOTKD} dengan <i>knowledge distillation</i> . Kalimat paralel diberikan sebagai <i>input</i> pada model guru dan model murid. vektor yang dihasilkan oleh model guru dan model murid di- <i>align</i> menggunakan fungsi <i>loss mean squared error</i>	56
Gambar 4.3.	Interpretasi dari model BERT _{CAT} dengan <i>integrated gradients</i> . Kata dengan warna hijau berarti kata tersebut berkontribusi positif terhadap hasil prediksi. Di lain sisi, kata yang berwarna merah berarti kata tersebut berkontribusi negatif terhadap hasil prediksi.	62

DAFTAR TABEL

Tabel 2.1.	Potongan <i>file</i> korpus <i>dataset</i> Miracl.	3
Tabel 2.2.	Potongan <i>file</i> kueri <i>dataset</i> Miracl.	3
Tabel 2.3.	Potongan <i>file</i> judgements <i>dataset</i> Miracl.	4
Tabel 2.4.	Beberapa fungsi aktivasi yang sering digunakan pada <i>multilayer perceptron</i>	17
Tabel 4.1.	Spesifikasi perangkat lunak yang digunakan pada penelitian ini.	48
Tabel 4.2.	Tabel Informasi untuk Setiap <i>Dataset</i> . Kolom <i>Korpus</i> menunjukkan jumlah entri pada <i>file korpus</i> , kolom <i>Kueri</i> menunjukkan jumlah entri pada <i>file kueri</i> , dan kolom <i>Judgements</i> menunjukkan jumlah entri pada <i>file judgements</i> (pasangan kueri dan teks dengan nilai relevansi).	50
Tabel 4.3.	Potongan <i>file</i> korpus <i>dataset</i> Miracl.	50
Tabel 4.4.	Potongan <i>file</i> kueri <i>dataset</i> Miracl.	51
Tabel 4.5.	Potongan <i>file</i> judgements <i>dataset</i> Miracl.	51
Tabel 4.6.	Potongan <i>dataset</i> yang digunakan untuk pelatihan model IndoBERT _{CAT}	52
Tabel 4.7.	<i>Hyperparameter</i> yang digunakan untuk <i>fine tuning</i> IndoBERT _{CAT}	53
Tabel 4.8.	<i>Hyperparameter</i> yang digunakan untuk <i>fine tuning</i> IndoBERT _{DOT}	54
Tabel 4.9.	Potongan <i>file</i> sentence-transformers/msmarco-hard-negatives. kolom <i>qid</i> berisikan id dari kueri, kolom <i>positive</i> adalah id teks positif, dan kolom <i>hard negative</i> adalah id teks yang sulit dibedakan dengan teks positif menggunakan BM25.	54
Tabel 4.10.	<i>Hyperparameter</i> yang digunakan untuk <i>fine tuning</i> IndoBERT _{DOTHardnegs}	55
Tabel 4.11.	Potongan dari <i>dataset</i> yang digunakan untuk pelatihan model IndoBERT _{KD}	56
Tabel 4.12.	<i>Hyperparameter</i> yang digunakan untuk <i>fine tuning</i> IndoBERT _{DOTKD}	57
Tabel 4.13.	Evaluasi model IndoBERT _{CAT} pada <i>dataset</i> Mmarco <i>dev set</i> , MrTyDi <i>test set</i> , dan Miracl <i>dev set</i>	57
Tabel 4.14.	Evaluasi model IndoBERT _{DOT} pada <i>dataset</i> Mmarco <i>dev set</i> , MrTyDi <i>test set</i> , dan Miracl <i>dev set</i>	58
Tabel 4.15.	Evaluasi model IndoBERT _{DOTHardnegs} pada <i>dataset</i> Mmarco <i>dev set</i> , MrTyDi <i>test set</i> , dan Miracl <i>dev set</i>	59
Tabel 4.16.	Evaluasi model IndoBERT _{KD} pada <i>dataset</i> Mmarco <i>dev set</i> , MrTyDi <i>test set</i> , dan Miracl <i>dev set</i>	59
Tabel 4.17.	Evaluasi dari model IndoBERT _{CAT} , IndoBERT _{DOT} , IndoBERT _{DOTHardnegs} , dan IndoBERT _{DOTKD} pada <i>dataset</i> Mmarco <i>dev set</i> , MrTyDi <i>test set</i> , dan Miracl <i>dev set</i>	60
Tabel 4.18.	<i>Benchmark</i> model BERT _{DOT} dan BERT _{CAT} , dan BM25 pada <i>dataset</i> Mmarco <i>dev set</i> . Latensi dan memori diukur pada <i>hardware</i> yang sama dengan yang digunakan pada pelatihan model.	61

Tabel 4.19. Interpretasi dari model BERT_{DOT} dengan menghitung hasil kali titik antara vektor teks dengan vektor masing-masing kata pada teks tersebut. Hanya 5 kata dengan nilai *importance* tertinggi yang ditunjukkan. 63

DAFTAR KODE PROGRAM

DAFTAR LAMPIRAN

Lampiran 1. CHANGELOG	69
Lampiran 2. Judul Lampiran 2	71

BAB 1

PENDAHULUAN

@todo

wew

BAB 2

LANDASAN TEORI

2.1 Masalah Pemeringkatan Teks

Permasalahan pemeringkatan teks adalah Permasalahan untuk menentukan urutan teks yang paling relevan dengan kueri q yang diberikan. Dalam bahasa yang lebih formal, 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 . Selain itu, biasanya nilai k akan lebih kecil dari banyaknya teks yang ada, sehingga permasalahan pemeringkatan sering juga disebut sebagai *top-k retrieval*. Untuk mengukur performa suatu model pemeringkatan, biasanya digunakan metrik evaluasi seperti presisi, *recall*, *reciprocal rank*, dan *normalized discounted cumulative gain* (nDCG) yang akan dijelaskan pada Subbab 2.1.2.

2.1.1 Bentuk Umum *Dataset*

Sebelum menjelaskan metrik evaluasi, akan dijelaskan terlebih dahulu bentuk umum dari *dataset* yang digunakan untuk mengevaluasi sebuah sistem pemeringkatan teks. Bentuk umum dari *dataset* yang digunakan biasanya terdiri dari 3 *file*, yaitu *file* korpus, *file* kueri, dan *file judgements*. *File* korpus adalah kumpulan teks yang ingin di-*retrieve* oleh sebuah sistem pemeringkatan teks. Biasanya, pada *file* korpus terdapat 3 kolom, yaitu id teks, judul teks, dan isi dari teks tersebut. Tabel 2.1 menunjukkan potongan dari *file* korpus.

Tabel 2.1: Potongan *file* korpus *dataset* Miracl.

_id	title	text
1342516#1	Colobothea biguttata	Larva kumbang ini biasanya mengebor ke dalam kayu dan dapat menyebabkan kerusakan pada batang kayu hidup atau kayu yang telah ditebang.
1342517#0	Ichthyodes rufipes	Ichthyodes rufipes adalah spesies kumbang tanduk panjang yang berasal dari famili Cerambycidae. Spesies ini juga merupakan bagian dari genus Ichthyodes, ordo Coleoptera, kelas Insecta, filum Arthropoda, dan kingdom Animalia.

file kueri berisi kumpulan kueri yang digunakan untuk mengambil teks dari *file* korpus. performa dari sistem pemeringkatan teks akan diukur dengan mengambil k teks dari *file* korpus untuk setiap kueri pada *file* kueri. Biasanya, pada *file* kueri terdapat 2 kolom, yaitu id kueri dan isi dari kueri tersebut. Tabel 2.2 menunjukkan potongan dari *file* kueri.

Tabel 2.2: Potongan *file* kueri *dataset* Miracl.

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

Selanjutnya *file judgements* berisi pemetaan relevansi antara kueri pada *file* kueri dengan teks pada *file* korpus. Biasanya, pada *file judgements* terdapat 3 kolom, yaitu id kueri, id teks, dan relevansi r antara kueri dan teks tersebut. Pasangan (kueri, teks) yang relevan akan memiliki nilai $r > 0$ dan nilai r yang makin besar menunjukkan relevansi yang makin tinggi. Selain itu, pasangan (kueri, teks) yang tidak relevan akan memiliki nilai $r = 0$ dan biasanya pasangan (kueri, teks) yang tidak relevan tidak dituliskan pada *file judgements*. tak menutup kemungkinan jika sebuah *dataset* hanya menggunakan nilai relevansi biner ($r \in \{0, 1\}$). Terakhir, Tabel 2.3 menunjukkan potongan dari *file judgements*.

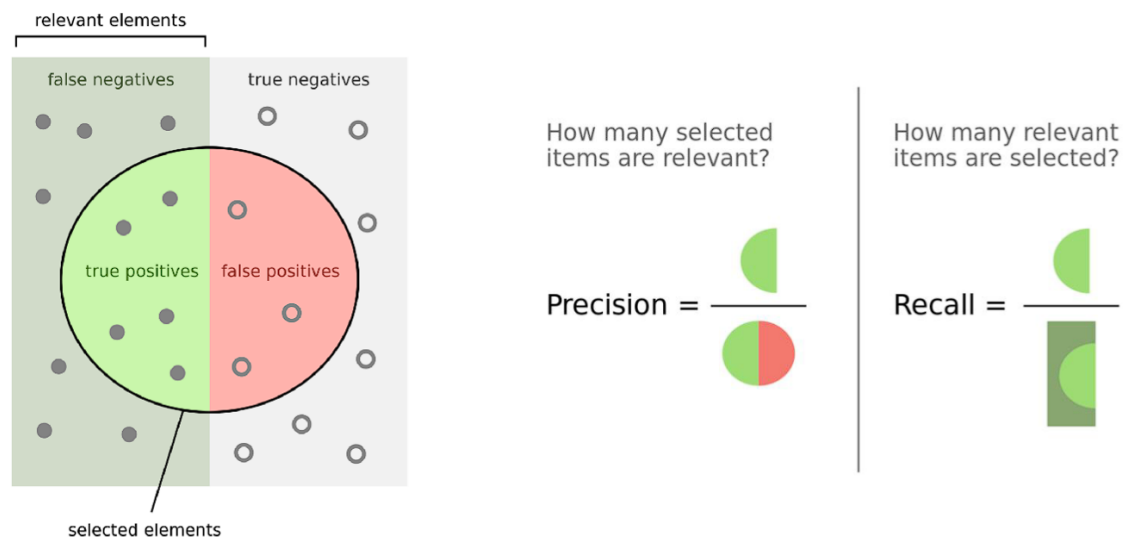
Tabel 2.3: Potongan *file judgements dataset* Miracl.

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

2.1.2 Metrik Evaluasi dalam Pemeringkatan Teks

Subbab ini menjelaskan beberapa metrik evaluasi yang sering digunakan untuk mengukur performa dari sistem pemeringkatan teks. Metrik evaluasi yang akan dijelaskan adalah *recall*, presisi, *reciprocal rank*, dan *normalized discounted cumulative gain* (nDCG). Metrik tersebut digunakan untuk mengukur performa dari sistem pemeringkatan teks dengan mengambil k teks dari *file* korpus pada satu kueri. Untuk mendapatkan performa dari sistem pemeringkatan teks secara keseluruhan, biasanya metrik evaluasi tersebut akan dihitung untuk setiap kueri pada *file* kueri dan kemudian diambil nilai rata-ratanya.

2.1.2.1 *Recall* dan Presisi



Gambar 2.1: Ilustrasi *recall* dan presisi. Nilai *recall* dihitung sebagai rasio teks relevan yang diambil oleh sistem terhadap seluruh teks yang relevan dengan kueri q . Sedangkan nilai presisi dihitung sebagai rasio teks relevan yang diambil oleh sistem terhadap seluruh teks yang diambil oleh sistem.

Sumber: (Hofstätter et al., 2021).

Presisi dan *recall* adalah metrik yang paling sederhana untuk mengukur kemampuan dari suatu sistem pemeringkatan teks. *Recall* mengukur kemampuan sistem dalam mengembalikan semua teks yang relevan dengan kueri q dari himpunan teks \mathcal{D} , sedangkan presisi mengukur kemampuan sistem dalam mengembalikan teks yang relevan dengan kueri q dari himpunan teks \mathcal{D} . Untuk suatu kueri q , kumpulan teks $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$, dan barisan k teks yang diambil oleh sistem, $D_k = (d_{i_1}, d_{i_2}, \dots, d_{i_k})$, *recall* dan presisi dapat dihitung dengan Persamaan 2.1 hingga Persamaan 2.2.

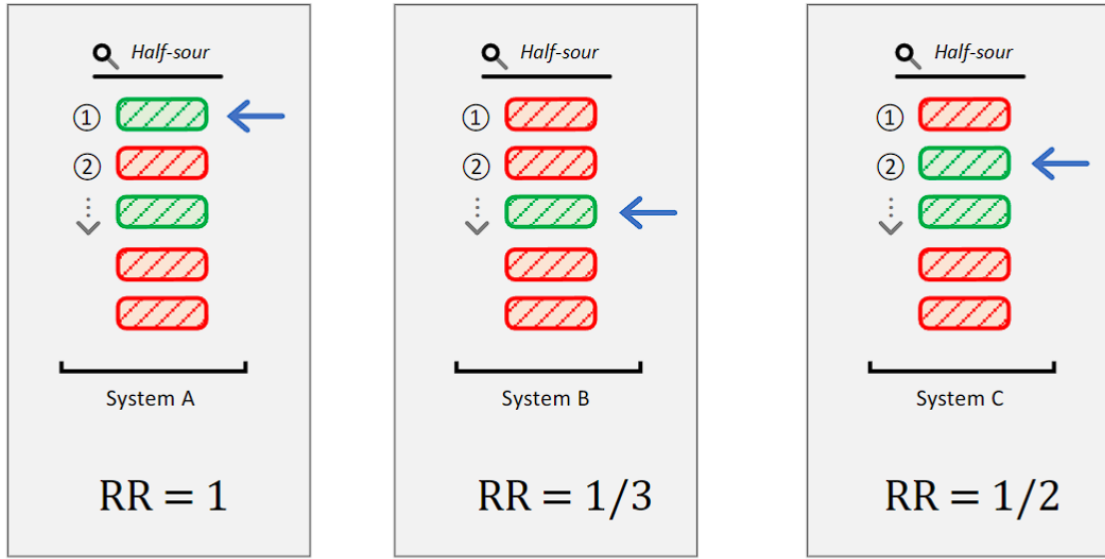
$$\text{recall}(q, D_k)@k = \frac{\sum_{d \in D_k} \text{rel}(q, d)}{\sum_{d \in \mathcal{D}} \text{rel}(q, d)} \in [0, 1], \quad (2.1)$$

$$\text{precision}(q, D_k)@k = \frac{\sum_{d \in D_k} \text{rel}(q, d)}{|D_k|} \in [0, 1], \quad (2.2)$$

$$\text{rel}(q, d) = \begin{cases} 1 & \text{jika } r > 1 \\ 0 & \text{jika } r = 0 \end{cases}. \quad (2.3)$$

Sebagai Contoh, Jika terdapat 10 teks yang relevan dengan kueri q , dan sistem mengembalikan $k = 100$ teks, namun hanya terdapat 5 teks yang relevan pada D_k maka *recall* dan presisi dari sistem tersebut adalah 0.5 ($\frac{5}{10}$) dan 0.05 ($\frac{5}{100}$) masing-masing. Baik *recall* maupun presisi memiliki rentang nilai dari 0 hingga 1, dengan nilai 1 menunjukkan performa sistem yang terbaik. perhitungan *recall* biasanya dilakukan untuk k yang cukup besar ($k = 100, 1000$), sedangkan perhitungan presisi dilakukan untuk k yang kecil ($k = 1, 3, 5$) (Hofstätter et al., 2021).

2.1.2.2 Reciprocal Rank



Gambar 2.2: Ilustrasi *reciprocal rank* (RR). Nilai RR pada sistem A, B, dan C berturut-turut adalah 1, 0.33, dan 0.5 karena posisi dari teks yang relevan pertama adalah 1, 3, dan 2.

Sumber: (Hofstätter et al., 2021).

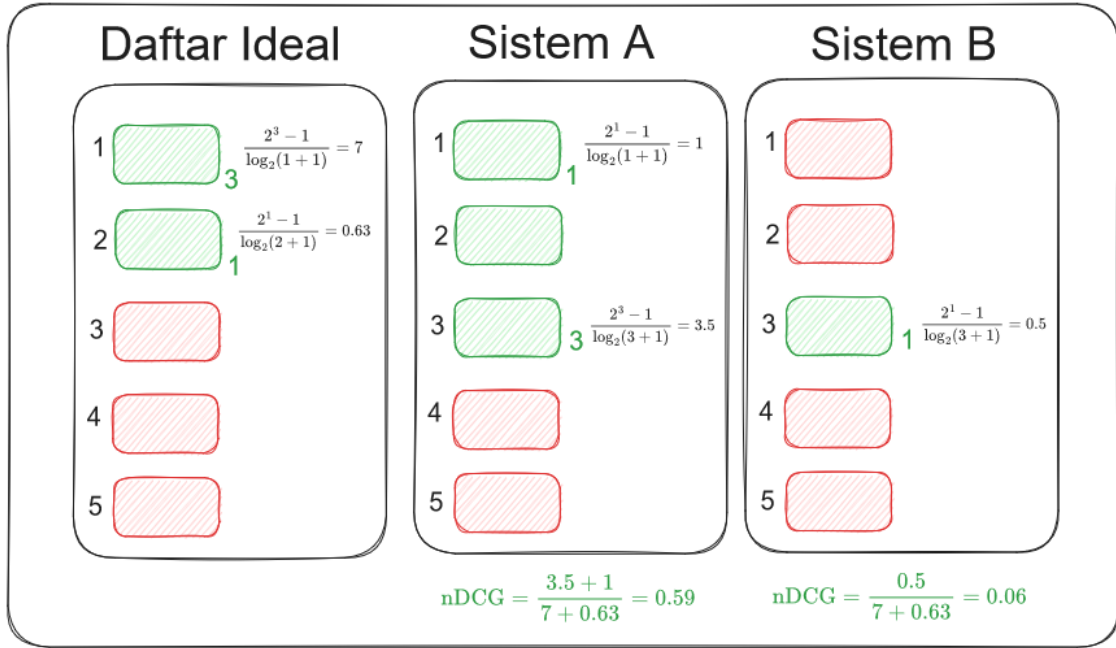
Metrik lainnya yang sering digunakan untuk mengukur performa sistem pemeringkatan adalah *reciprocal rank* (RR). Metrik RR menitikberatkan pada peringkat dari teks relevan pertama dengan kueri q . Persamaan 2.4 hingga Persamaan 2.5 menunjukkan cara menghitung RR dari suatu kueri q dan barisan k teks yang diambil oleh sistem.

$$RR(q, D_k)@k = \begin{cases} \frac{1}{\text{FirstRank}(q, D_k)} & \text{jika } \exists d \in D_k \text{ dengan } \text{rel}(q, d) = 1 \\ 0 & \text{jika } \forall d \in D_k, \text{rel}(q, d) = 0 \end{cases} \in [0, 1], \quad (2.4)$$

$$\text{FirstRank}(q, D_k) = \text{posisi teks relevan pertama } d \in D_k \text{ dengan } \text{rel}(q, d) = 1. \quad (2.5)$$

Gambar 2.2 mengilustrasikan metrik RR. Pada gambar tersebut, nilai RR dari sistem A adalah 1 ($\frac{1}{1}$) karena posisi dari teks yang relevan pertama adalah 1. Nilai RR dari sistem B dan sistem C masing-masing adalah 0.33 ($\frac{1}{3}$) dan 0.5 ($\frac{1}{2}$) karena posisi dari teks yang relevan pertama adalah 3 dan 2. Selain itu, jika tidak terdapat teks yang relevan dengan kueri q pada D_k , nilai RR dari sistem tersebut adalah 0.

2.1.2.3 Normalized Discounted Cumulative Gain (nDCG)



Gambar 2.3: Ilustrasi perhitungan nDCG. Nilai nDCG dari sistem A, B, adalah rasio antara DCG dari sistem tersebut dengan DCG dari sistem ideal.

Sumber: (Hofstätter et al., 2021), telah diolah kembali.

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 *judgements* r yang tak biner. Fungsi $\text{rel}(q, d)$ pada Persamaan 2.3 berubah menjadi $\text{rel}(q, d) = r$ ketika menghitung metrik nDCG. Persamaan 2.6 hingga Persamaan 2.8 menunjukkan cara menghitung nDCG dari suatu kueri q dan barisan k teks yang diambil oleh sistem.

$$\text{nDCG}(q, D_k)@k = \frac{\text{DCG}(q, D_k)@k}{\text{DCG}(q, D_k^{\text{ideal}})@k} \in [0, 1], \quad (2.6)$$

$$\text{DCG}(q, D_k)@k = \sum_{d \in D_k} \frac{2^{\text{rel}(q, d)} - 1}{\log_2(\text{rank}(d, D_k) + 1)}, \quad (2.7)$$

$$\text{rank}(d, D_k) = \text{Posisi } d \text{ dalam } D_k. \quad (2.8)$$

Perhitungan *discounted cumulative gain* (DCG) pada Persamaan 2.7 dapat dijelaskan menjadi dua faktor berikut:

1. faktor $2^{\text{rel}(q, d)} - 1$ menunjukkan bahwa teks yang lebih relevan akan memiliki nilai

yang lebih tinggi dari teks yang kurang relevan untuk posisi teks yang sama.

2. faktor $\frac{1}{\log_2(\text{rank}(d, D_k) + 1)}$ menunjukkan bahwa teks yang relevan yang muncul pada peringkat yang lebih tinggi akan memiliki nilai yang lebih tinggi dari teks dengan relevansi yang sama, tetapi muncul pada peringkat yang lebih rendah.

nilai dari nDCG pada Persamaan 2.6 adalah nilai DCG pada barisan teks D_k yang dinormalisasi oleh nilai DCG pada barisan teks ideal D_k^{ideal} . Barisan teks ideal D_k^{ideal} adalah barisan teks yang diurutkan berdasarkan relevansinya dengan kueri q .

Selain itu, jika pada *dataset* memiliki *judgements* biner, faktor $2^{\text{rel}(q, d)} - 1$ pada Persamaan 2.7 dapat diubah menjadi $\text{rel}(q, d)$. Akibatnya, Persamaan 2.7 akan menjadi Persamaan 2.9.

$$\text{DCG}(q, D_k)@k = \sum_{d \in D_k} \frac{\text{rel}(q, d)}{\log_2(\text{rank}(d, D_k) + 1)}. \quad (2.9)$$

2.2 Pemeringkatan Teks dengan Statistik

Untuk mengambil k teks dari kumpulan \mathcal{D} diperlukan suatu fungsi skor $s(q, d, \mathcal{D})$ yang mengukur relevansi antara kueri q dan teks d . dengan mencari skor antara q terhadap semua teks pada \mathcal{D} , Barisan teks $D_k = (d_{i_1}, d_{i_2}, \dots, d_{i_k})$ dapat dipilih sehingga $\text{score}(q, d_{i_1}) \geq \text{score}(q, d_{i_2}) \geq \dots \geq \text{score}(q, d_{i_k})$ adalah k teks dengan skor tertinggi.

Bagian ini menjelaskan dua fungsi skor statistik sederhana yang menjadi *baseline* ketika membandingkan performa dari model pemeringkatan teks yang lebih kompleks. Subbab 2.2.1 menjelaskan fungsi skor statistik yang berdasarkan pada frekuensi kemunculan kata pada teks dan kueri. Selanjutnya, Subbab 2.2.2 membahas fungsi skor statistik yang menjadi standar *de facto* dalam pemeringkatan teks.

2.2.1 Term Frequency - Inverse Document Frequency (TF-IDF)

fungsi skor TF-IDF adalah fungsi skor statistik yang menghitung score antara kueri q dan teks d dengan menghitung frekuensi kemunculan kata pada teks dan kueri. Untuk suatu kueri q , misalkan $T_q = \{t_1, t_2, \dots, t_{L_1}\}$ adalah himpunan kata yang terdapat pada kueri q . Selain itu, misalkan juga $T_d = \{t_1, t_2, \dots, t_n\}$ adalah himpunan kata yang terdapat pada teks d . Nilai skor antara q dan d diberikan oleh persamaan Persamaan 2.10 sampai

Persamaan 2.15.

$$\text{score}(q, d, \mathcal{D}) = \sum_{t \in T_q \cap T_d} \text{TF-IDF}(t, d, \mathcal{D}) \quad (2.10)$$

$$\text{TF-IDF}(t, d, \mathcal{D}) = \text{tf}(t, d) \times \text{idf}(t, \mathcal{D}) \quad (2.11)$$

$$\text{tf}(t, d) = \frac{\text{Count}(t, d)}{|d|} \quad (2.12)$$

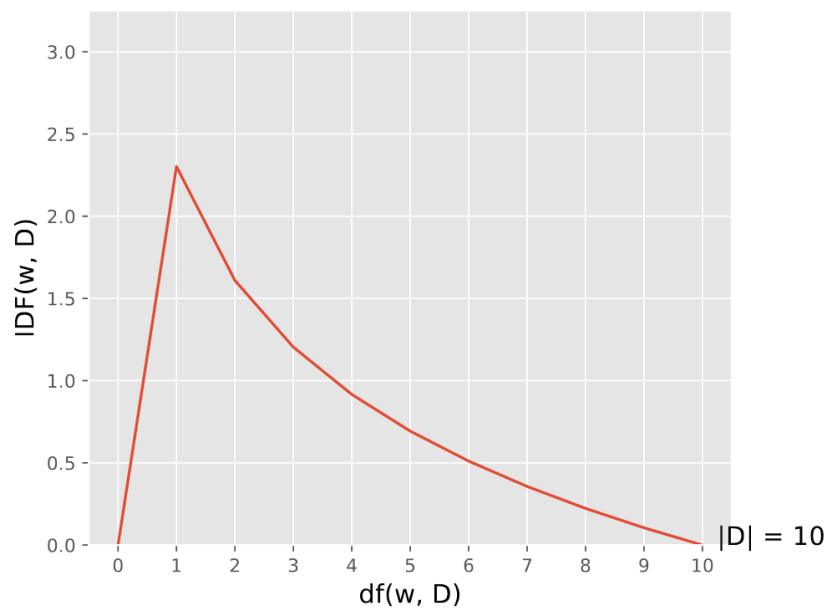
$$\text{Count}(t, d) = \text{jumlah kemunculan } t \text{ dalam } d \quad (2.13)$$

$$\text{idf}(t, \mathcal{D}) = \begin{cases} \log_2 \left(\frac{|\mathcal{D}|}{\text{df}(t, \mathcal{D})} \right) & \text{jika } \text{df}(t, \mathcal{D}) > 0 \\ 0 & \text{jika } \text{df}(t, \mathcal{D}) = 0 \end{cases} \quad (2.14)$$

$$\text{df}(t, \mathcal{D}) = \text{jumlah teks pada } \mathcal{D} \text{ yang mengandung } t \quad (2.15)$$

skor untuk pasangan (q, d) dihitung dengan menjumlahkan skor TF-IDF dari setiap kata yang terdapat pada kueri q dan teks d . skor TF-IDF dari suatu kata t adalah perkalian antara *term frequency* $\text{tf}(q, d)$ dan *inverse document frequency* $\text{idf}(t, \mathcal{D})$. Fungsi skor pada Persamaan 2.10 dapat dijelaskan menjadi dua faktor utama sebagai berikut:

1. faktor $\text{tf}(t, d)$ menunjukkan bahwa nilai TF-IDF meningkat seiring dengan bertambahnya frekuensi kemunculan kata t pada teks d .
2. Faktor $\text{idf}(t, \mathcal{D})$ menunjukkan bahwa nilai TF-IDF meningkat seiring dengan *rarity* dari kata t pada himpunan teks \mathcal{D} . Akibatnya, kata yang jarang muncul pada himpunan teks \mathcal{D} dan muncul pada suatu teks tertentu akan menghasilkan skor yang tinggi. Sementara itu, kata-kata yang sering muncul pada koleksi teks \mathcal{D} memiliki nilai *downgraded*. Gambar 2.4 menunjukkan grafik dari fungsi idf .



Gambar 2.4: Grafik dari fungsi idf. Nilai idf menurun seiring dengan bertambahnya nilai $df(t, \mathcal{D})$.
Sumber: (Potts et al., 2023).

Kata-kata seperti preposisi atau kata ganti akan menghasilkan skor TF-IDF yang sangat rendah. Hal ini menyiratkan bahwa kata-kata tersebut memiliki sedikit relevansi dalam teks dan bisa diabaikan. Di sisi lain, kata-kata yang muncul secara berlebihan dalam satu teks tetapi jarang muncul dalam teks lainnya akan menghasilkan nilai $tf(t, d)$ dan $\log\left(\frac{\mathcal{D}}{df(t, \mathcal{D})}\right)$ yang relatif besar. Dampaknya adalah skor TF-IDF yang dihasilkan juga menjadi signifikan. Gambar 2.5 menunjukkan contoh perhitungan skor TF-IDF untuk suatu kumpulan teks.

	doc ₁	doc ₂	doc ₃	doc ₄
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 ₁	doc ₂	doc ₃	doc ₄
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 ₁	doc ₂	doc ₃	doc ₄
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

Gambar 2.5: Ilustrasi perhitungan skor TF-IDF untuk suatu kumpulan teks.

Sumber: (Potts et al., 2023).

2.2.2 Best Match 25 (BM25)

BM25 (*Best Match attempt 25*) merupakan pengembangan dari fungsi skor TF-IDF dengan perbedaan utama pada fungsi nilai yang berkaitan dengan frekuensi kata – digunakan $\text{score}_{\text{BM25}}(q, d)$ (Persamaan 2.17) daripada $\text{tf}(q, d)$ (Persamaan 2.11). Pada fungsi $\text{score}_{\text{BM25}}(q, d)$ terdapat 2 parameter yang dapat diatur, yaitu b , dan k_1 . Setiap parameter mempunyai efek yang berbeda terhadap nilai $\text{score}_{\text{BM25}}(q, d)$ yang dihasilkan. Sebelum menjelaskan efek dari setiap parameter, Persamaan 2.16 hingga Persamaan 2.20 menunjukkan cara menghitung skor relevansi dari suatu kueri q dan teks d .

$$\text{score}(q, d, \mathcal{D}) = \sum_{t \in T_q \cap T_d} \text{BM25}(t, d, \mathcal{D}), \quad (2.16)$$

$$\text{BM25}(t, d, \mathcal{D}) = \text{idf}_{\text{BM25}}(t, \mathcal{D}) \times \text{score}_{\text{BM25}}(q, d, \mathcal{D}), \quad (2.17)$$

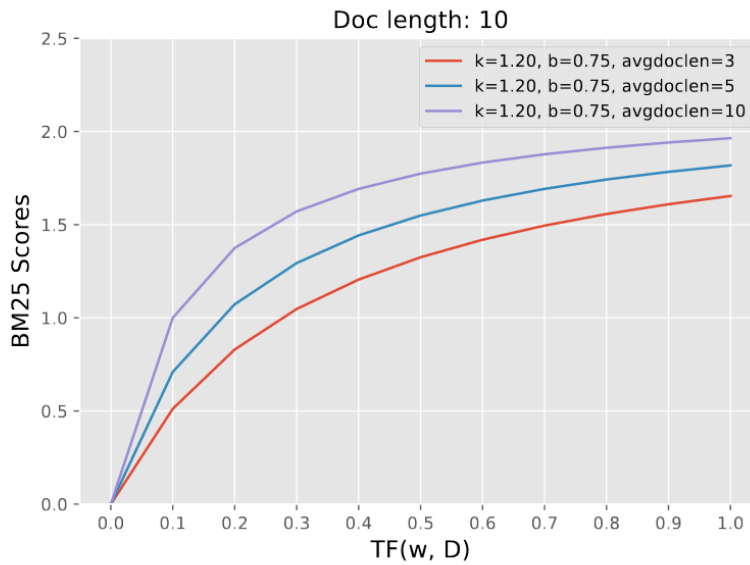
$$\text{score}_{\text{BM25}}(t, d) = \frac{\text{tf}(t, d) \times (k_1 + 1)}{\text{tf}(t, d) + k_1 \times (1 - b + b \times \frac{|d|}{\text{avgdl}})}, \quad (2.18)$$

$$\text{idf}_{\text{BM25}}(t, \mathcal{D}) = \log \left(1 + \frac{|\mathcal{D}| - \text{df}(t, \mathcal{D}) + 0.5}{\text{df}(t, \mathcal{D}) + 0.5} \right), \quad (2.19)$$

$$\text{avgdl} = \text{rata-rata panjang teks pada koleksi } \mathcal{D}. \quad (2.20)$$

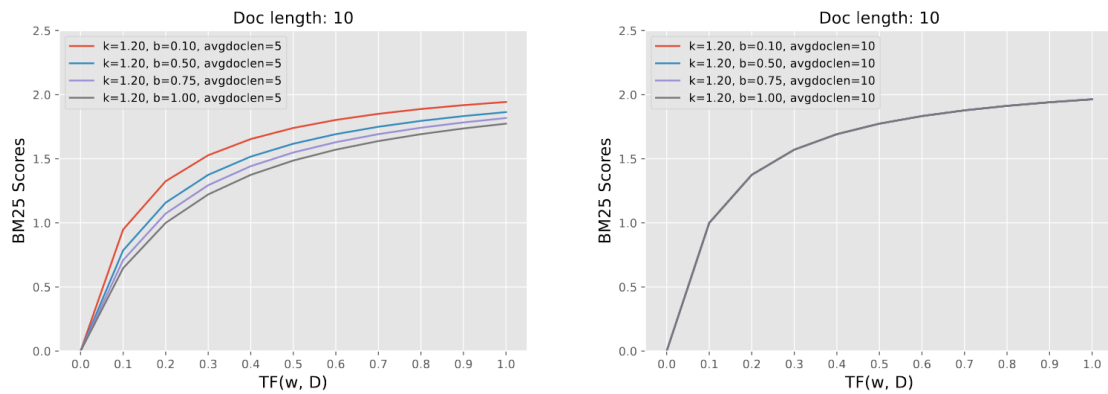
Efek dari masing-masing parameter dan faktor pada $\text{score}_{\text{BM25}}(t, d)$ dapat dijelaskan sebagai berikut:

1. Faktor $\frac{|d|}{\text{avgdl}}$ pada $\frac{\text{tf}(t,d) \times (k_1+1)}{\text{tf}(t,d) + k_1 \times \left(1 - b + b \times \frac{|d|}{\text{avgdl}}\right)}$ men-*penalize* skor pada teks yang panjangnya lebih besar dari rata-rata panjang teks pada himpunan teks \mathcal{D} . Gambar 2.6 Menunjukkan efek dari perbedaan nilai avgdl terhadap skor yang dihasilkan, makin besar rasio $\frac{|d|}{\text{avgdl}}$ makin kecil skor yang dihasilkan.
2. Nilai b menentukan seberapa besar efek dari faktor $\frac{|d|}{\text{avgdl}}$ terhadap skor yang dihasilkan. Gambar 2.7 Menunjukkan efek dari perbedaan nilai b terhadap skor yang dihasilkan. Untuk $\frac{|d|}{\text{avgdl}} = 1$, Faktor b tidak memiliki pengaruh terhadap skor. Nilai b yang umum dipilih berada pada rentang $[0.5, 0.8]$.
3. Nilai k_1 men-*penalize* kemunculan kata t pada teks d yang berlebih. Gambar 2.8 Menunjukkan efek dari perbedaan nilai k_1 terhadap skor yang dihasilkan. untuk nilai k_1 yang ekstrim, nilai $\text{score}_{\text{BM25}}(t, d)$ hanya menjadi indikator saja dari kemunculan kata t pada teks d . Nilai k_1 yang umum dipilih berada pada rentang $[1.2, 2.0]$.

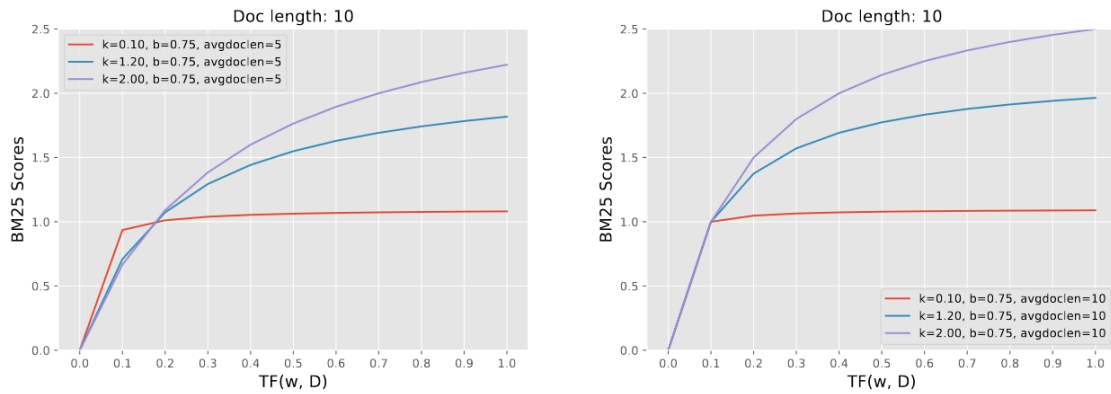


Gambar 2.6: Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai avgdl.

Sumber: (Potts et al., 2023).

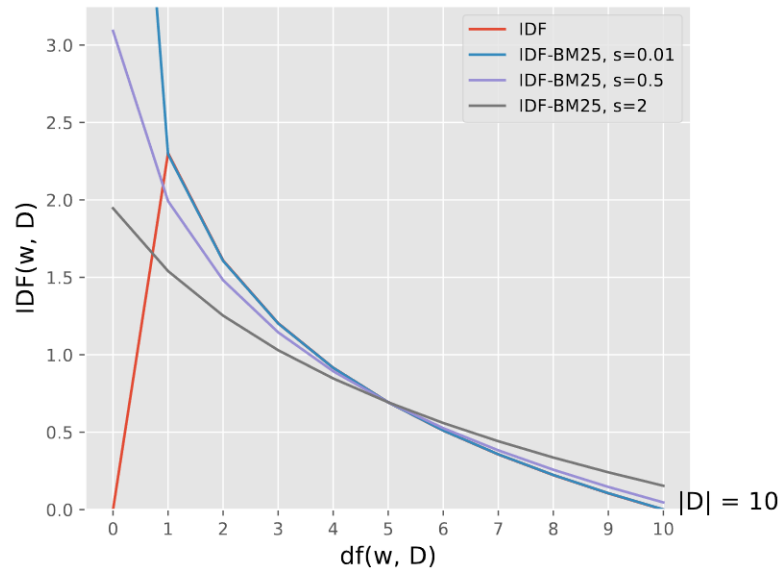


Gambar 2.7: Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai b .
Sumber: (Potts et al., 2023).



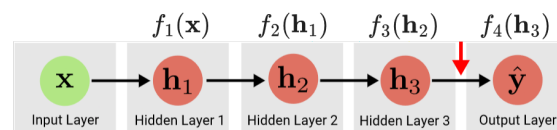
Gambar 2.8: Kumpulan grafik dari fungsi $\text{score}_{\text{BM25}}(t, d)$ dengan perbedaan nilai k_1 .
Sumber: (Potts et al., 2023).

Perbedaan *minor* lainnya ada pada fungsi idf. Fungsi idf pada BM25 merupakan versi *smoothing* dari idf dengan tujuan untuk menghindari nilai idf yang bernilai 0 ketika kata t tidak muncul pada himpunan teks \mathcal{D} – semata-mata untuk konsistensi dengan asumsi bahwa kata t yang tidak muncul pada himpunan teks \mathcal{D} memiliki nilai idf yang paling tinggi. Gambar 2.9 Menunjukkan Perbedaan antara idf_{BM25} dan idf. Perbedaan utamanya terjadi ketika $\text{df}(t, \mathcal{D}) = 0$, nilai dari idf_{BM25} tak nol dan mengikuti pola yang diharapkan. Ketika $\text{df}(t, \mathcal{D}) > 0$, nilai dari idf_{BM25} dan idf hampir serupa.



Gambar 2.9: Perbedaan antara idf_{BM25} dan idf .
Sumber: (Potts et al., 2023).

2.3 Deep Learning



Gambar 2.10: Ilustrasi dari *Directed Acyclic Graph* (DAG) pada arsitektur *deep learning feed-forward neural network* (FFN).

Sumber: (Geiger et al., 2022), telah diolah kembali.

Arsitektur *Deep learning* merujuk pada model *machine learning* yang tersusun dari fungsi-fungsi terturunkan (yang biasa disebut sebagai *layer*), dimana komposisi antara fungsi-fungsi tersebut dapat digambarkan sebagai *directed acyclic graph* (DAG) yang memetakan suatu *input* ke suatu *output*. Biasanya, setiap fungsi dalam Arsitektur *Deep learning* memiliki parameter yang ingin diestimasi atau dicari dengan data.

Gambar 2.10 menunjukkan arsitektur *deep learning* yang sederhana, yaitu *feed-forward neural network* (FFN). Pada Gambar 2.10, *input* x akan dipetakan ke *output* \hat{y} melalui serangkaian fungsi f_1, f_2, f_3, f_4 yang disebut sebagai *layer*. Setiap *layer* f_i memiliki parameter θ_i yang akan diestimasi dengan data. Selain itu, *Output* dari *layer* f_i akan menjadi *input* dari *layer* f_{i+1} . *Output* dari *layer* f_4 adalah *output* dari model. Model pada

Gambar 2.10 dapat ditulis sebagai Persamaan 2.21.

$$\hat{y} = f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) = f_4(f_3(f_2(f_1(\mathbf{x}; \boldsymbol{\theta}_1); \boldsymbol{\theta}_2); \boldsymbol{\theta}_3); \boldsymbol{\theta}_4), \quad (2.21)$$

$$(2.22)$$

dengan $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4\}$ adalah parameter dari model.

2.3.1 *Multilayer Perceptron (MLP)*

Multi-layer perceptron (MLP) adalah *feed-forward neural network* dengan setiap fungsi f_i adalah fungsi linear yang diikuti oleh fungsi aktivasi non-linear ϕ yang diterapkan *element-wise* pada setiap *output*-nya. *Hyperparameter* lainnya selain fungsi aktivasi adalah kedalaman model L , dan dimensi *output* dari setiap *layer* d_1, d_2, \dots, d_L .

Untuk permasalahan regresi dengan *input* $\mathbf{x} \in \mathbb{R}^{d_0}$ dan *output* $\mathbf{y} \in \mathbb{R}^{d_L}$, Persamaan 2.23 hingga Persamaan 2.25 menunjukkan arsitektur MLP untuk permasalahan regresi dengan L *layer* dan fungsi aktivasi ϕ .

$$f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) = f_L(f_{L-1}(\dots f_1(\mathbf{x})) \dots), \quad (2.23)$$

$$f_L(\mathbf{x}) = \mathbf{x}\mathbf{W}_L + \mathbf{b}_L \in \mathbb{R}^{d_L}, \quad (2.24)$$

$$f_l(\mathbf{x}; \mathbf{W}_l, \mathbf{b}_l) = \phi(\mathbf{x}\mathbf{W}_l + \mathbf{b}_l) \in \mathbb{R}^{d_l}, \quad l = 1, 2, \dots, L-1, \quad (2.25)$$

dengan keterangan sebagai berikut:

$\phi(\mathbf{x})$ = fungsi aktivasi non-linear,

$\boldsymbol{\theta} = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, \dots, \mathbf{W}_L, \mathbf{b}_L\}$,

\mathbf{W}_l = matriks bobot $\in \mathbb{R}^{d_{l-1} \times d_l}$,

\mathbf{b}_l = vektor bias $\in \mathbb{R}^{d_l}$.

Untuk Permasalahan Klasifikasi Biner dengan *input* $\mathbf{x} \in \mathbb{R}^{d_0}$ dan *output* $\mathbf{y} \in \{0, 1\}$, Persamaan 2.26 hingga Persamaan 2.30 menunjukkan arsitektur MLP untuk permasa-

lahan klasifikasi biner.

$$f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) = f_L(f_{L-1}(\dots f_1(\mathbf{x})) \dots), \quad (2.26)$$

$$f_L(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{x}\mathbf{W}_L + \mathbf{b}_L \in \mathbb{R}), \quad (2.27)$$

$$\boldsymbol{\sigma}(x) = \frac{1}{1 + e^x} \in (0, 1), \quad (2.28)$$

$$\text{decision}(\mathbf{x}; \boldsymbol{\theta}) = \begin{cases} 1 & \text{jika } f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) \geq \text{threshold} \\ 0 & \text{jika } f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) < \text{threshold} \end{cases}, \quad (2.29)$$

$$\text{threshold} \in [0, 1]. \quad (2.30)$$

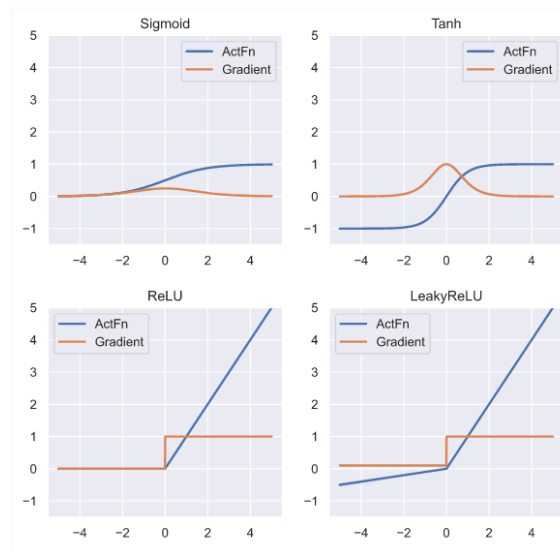
Perbedaan utama antara MLP untuk permasalahan regresi dan klasifikasi adalah fungsi aktivasi pada *output layer*. Pada permasalahan regresi, fungsi aktivasi pada *output layer* adalah fungsi identitas, sedangkan pada permasalahan klasifikasi, fungsi aktivasi pada *output layer* adalah fungsi *sigmoid*. Tujuan penggunaan fungsi *sigmoid* pada permasalahan klasifikasi adalah untuk memastikan bahwa *output* dari model berada pada rentang $[0, 1]$, nilai tersebut dapat diinterpretasikan sebagai probabilitas \mathbf{x} termasuk pada kelas positif. Selain itu, *threshold* pada Persamaan 2.30 digunakan untuk menentukan kelas dari \mathbf{x} .

2.3.2 Fungsi Aktivasi

Fungsi Aktivasi pada setiap fungsi f_i pada *multilayer perceptron* digunakan untuk menambahkan non-linearitas pada model. Sebab, tanpa adanya fungsi aktivasi non-linear, model *multilayer perceptron* akan menjadi model linear. Selain itu, fungsi aktivasi juga biasanya adalah fungsi yang terturunkan, meskipun tidak perlu terturunkan disetiap titik. Tabel 2.4 menunjukkan beberapa fungsi aktivasi yang sering digunakan pada *multilayer perceptron*. Gambar 2.11 menunjukkan grafik dari fungsi aktivasi pada Tabel 2.4 dan turunannya.

Tabel 2.4: Beberapa fungsi aktivasi yang sering digunakan pada *multilayer perceptron*.

Fungsi Aktivasi	Persamaan
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$
Tanh	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
ReLU	$\text{ReLU}(x) = \max(0, x)$
Leaky ReLU	$\text{LeakyReLU}(x) = \max(\alpha x, x), \alpha \in [0, 1]$

**Gambar 2.11:** Grafik dari fungsi aktivasi pada Tabel 2.4 dan turunannya.**Sumber:** (Lippe, 2022).

2.3.3 Fungsi Loss

Misalkan $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ adalah *dataset* yang terdiri dari n pasangan *input* dan *output*. Parameter θ pada f_{model} diestimasi dengan melakukan *fitting* pada *dataset* \mathcal{D} . Untuk melakukan *fitting* pada *dataset* \mathcal{D} , diperlukan suatu fungsi *loss* yang mengukur seberapa baik hasil pemetaan f_{model} pada *input* \mathbf{x}_i terhadap *output* \mathbf{y}_i . Meskipun sembarang fungsi yang terturunkan dapat digunakan sebagai fungsi *loss*, namun pemilihan fungsi *loss* berdasarkan *maximum likelihood estimation* (MLE) lebih disarankan.

Untuk permasalahan klasifikasi biner, fungsi *loss* yang sering digunakan adalah *binary cross entropy* (BCE) seperti yang ditunjukkan pada Persamaan 2.39. Penurunan fungsi *loss* BCE dengan mengikuti prinsip MLE yang akan dijelaskan pada bagian berikut.

Misalkan $y_i | \mathbf{x}$ mengikuti distribusi bernouli dengan parameter $p = f_{\text{model}}(\mathbf{x}; \theta)$ yang saling independen antara satu sama lainnya. Persamaan 2.31 menunjukkan definisi dari

$y_i \mid \mathbf{x}$.

$$y_i \mid \mathbf{x} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})), \quad (2.31)$$

$$p(y_i \mid \mathbf{x}) = f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})^{y_i} (1 - f_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}))^{1-y_i}. \quad (2.32)$$

Fungsi *likelihood* dari $\boldsymbol{\theta}$ terhadap *dataset* \mathcal{D} dapat ditulis sebagai berikut:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^N p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}). \quad (2.33)$$

Dengan prinsip MLE, parameter $\boldsymbol{\theta}$ yang dicari adalah parameter $\boldsymbol{\theta}$ yang memaksimalkan fungsi *likelihood* $\mathcal{L}(\boldsymbol{\theta})$,

$$\boldsymbol{\theta}_{\text{MLE}} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}). \quad (2.34)$$

Untuk mempermudah perhitungan, fungsi *likelihood* diubah menjadi negatif *log-likelihood* $\ell(\boldsymbol{\theta})$, sehingga permasalahan optimasi dapat ditulis seperti Persamaan 2.35 hingga Persamaan 2.37.

$$\ell(\boldsymbol{\theta}) = -\log \mathcal{L}(\boldsymbol{\theta}), \quad (2.35)$$

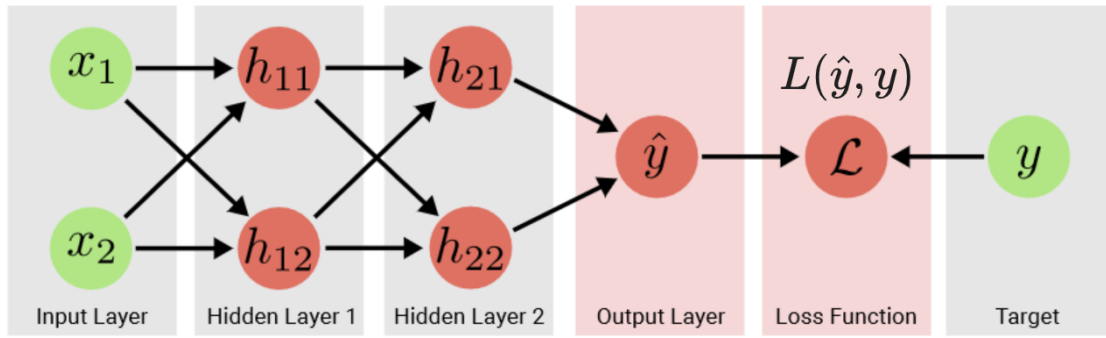
$$= -\sum_{i=1}^N \log(p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})), \quad (2.36)$$

$$\boldsymbol{\theta}_{\text{MLE}} = \arg \min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}). \quad (2.37)$$

Dengan mengganti $p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ dengan fungsi distribusi-nya, maka fungsi *loss* yang digunakan untuk permasalahan klasifikasi biner adalah *binary cross entropy* (BCE) seperti pada Persamaan 2.39. Gambar 2.12 mengilustrasikan *directed acyclic graph* (DAG) dari model ketika proses pelatihan dilakukan.

$$\boldsymbol{\theta}_{\text{MLE}} = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^N \underbrace{-y_i \log(f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta})) - (1 - y_i) \log(1 - f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta}))}_{\text{Binary Cross Entropy Loss } L(y_i, f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta}))}, \quad (2.38)$$

$$L(y_i, f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta})) = -y_i \log(f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta})) - (1 - y_i) \log(1 - f_{\text{model}}(\mathbf{x}_i; \boldsymbol{\theta})). \quad (2.39)$$



Gambar 2.12: Ilustrasi dari *Directed Acyclic Graph* (DAG) pada model *deep learning* ketika proses pelatihan dilakukan.

Sumber: (Geiger et al., 2022), telah diolah kembali.

Untuk mendapatkan f_{model} dengan performa yang baik, dibutuhkan model dengan nilai $\ell(\theta)$ semimumimum mungkin. Namun, pencarian θ sehingga $\ell(\theta)$ minimum secara analitik tidak dapat dilakukan karena non-linearitas yang ada pada model, dengan kata lain solusi dari $\nabla_{\theta}\ell(\theta) = 0$ tidak dapat dicari secara analitik. Sebagai gantinya, pencarian θ dilakukan secara numerik dengan menggunakan metode *gradient descent* yang akan dijelaskan pada bagian selanjutnya.

2.3.4 Optimasi Parameter

Gradient Descent adalah metode numerik yang digunakan untuk mencari nilai θ yang meminimalkan fungsi *loss* $\ell(\theta)$. Pada metode *gradient descent*, nilai θ diupdate secara iteratif dengan mengikuti arah negatif dari *gradient* $\nabla_{\theta}\ell(\theta)$ yang menunjukkan arah dari penurunan fungsi *loss* $\ell(\theta)$. Untuk kumpulan data $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, Persamaan 2.40 menunjukkan algoritma *gradient descent* untuk mencari nilai θ .

$$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} L(y_i, f_{\text{model}}(\mathbf{x}_i; \theta^{(t)})), \quad (2.40)$$

dengan $\eta \in \mathbb{R}^+$ adalah *learning rate* yang menentukan seberapa besar perubahan pada θ pada setiap iterasi.

Perlu diketahui bahwa pada metode *gradient descent* memperbarui parameter dengan mengambil rata-rata *gradient* dari semua data pada *dataset* pelatihan \mathcal{D} . Hal ini menciptakan masalah ketika model menggunakan banyak parameter dan jumlah data pada *datasets* latih besar, yaitu komputasi *forward pass* dan *backward pass* menjadi sangat

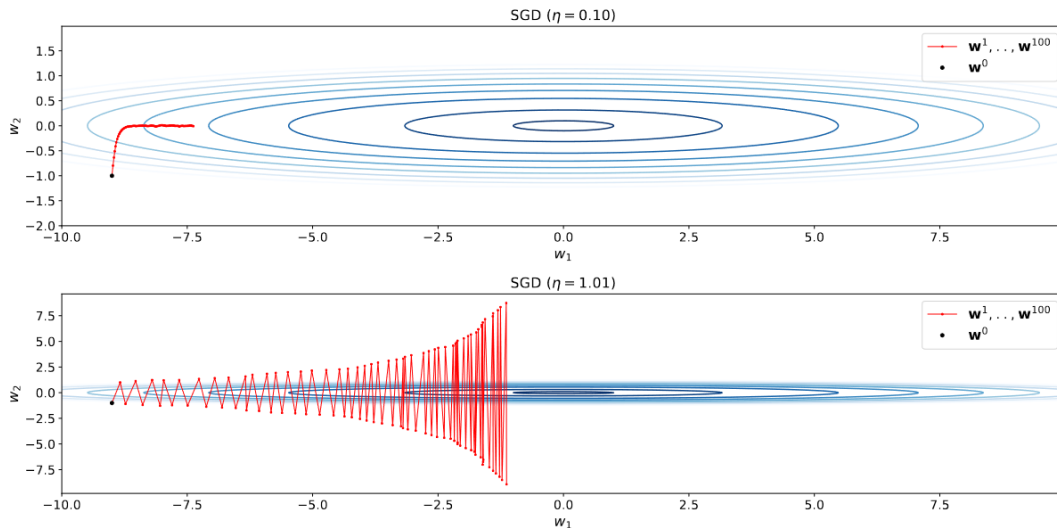
mahal dan diperlukan memori yang besar untuk menyimpan gradient dari semua data pada *dataset* latih. Untuk mengatasi masalah tersebut, digunakan metode *stochastic gradient descent* (SGD) dimana setiap *update* dari parameter θ dihitung dengan mengambil rata-rata *gradient* dari sebagian data pada *dataset* $\mathcal{B} \subseteq \mathcal{D}$. Persamaan 2.43 menunjukkan algoritma *stochastic gradient descent*.

$$\mathcal{B} = \{(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \dots, (\mathbf{x}_{i_b}, y_{i_b})\} \subseteq \mathcal{D}, |\mathcal{B}| \ll |\mathcal{D}|, \quad (2.41)$$

$$\nabla_{\theta} \mathcal{L}_{\mathcal{B}}(\theta) = \frac{1}{b} \sum_{i=1}^b \nabla_{\theta} L(y_i, f_{\text{model}}(\mathbf{x}_i; \theta)), \quad (2.42)$$

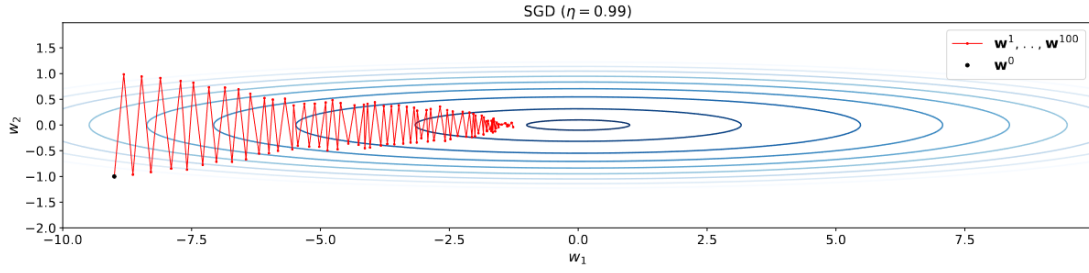
$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} \mathcal{L}_{\mathcal{B}}(\theta^{(t)}). \quad (2.43)$$

hyperparameter learning rate mengatur laju dari perubahan parameter θ pada setiap iterasi pembaruan. Dengan demikian, pemilihan *learning rate* berpengaruh terhadap kekonvergenan optimasi yang dilakukan. Jika *learning rate* yang digunakan terlalu kecil, model membutuhkan waktu yang jauh lebih lama untuk mencapai nilai parameter θ yang optimal. Di lain sisi, pemilihan *learning rate* yang terlalu besar dapat membuat model tidak dapat menemukan nilai parameter θ yang optimal. Gambar 2.13 mengilustrasikan proses pembaruan parameter θ dengan *learning rate* yang terlalu kecil dan terlalu besar, dan Gambar 2.14 mengilustrasikan proses pembaruan parameter θ dengan *learning rate* yang baik.



Gambar 2.13: 100 iterasi pertama dari pembaruan parameter $\theta = \{w_1, w_2\}$ dengan *learning rate* yang terlalu kecil dan terlalu besar.

Sumber: (Geiger et al., 2022).



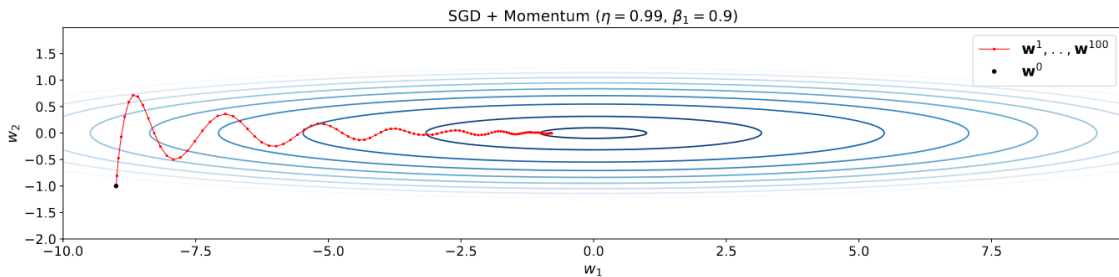
Gambar 2.14: 100 iterasi pertama dari pembaruan parameter $\theta = \{w_1, w_2\}$ dengan *learning rate* yang baik.

Sumber: (Geiger et al., 2022).

Untuk mempercepat proses pembaruan parameter θ , digunakan metode *stochastic gradient descent* dengan momentum untuk mengurangi osilasi pada proses pembaruan parameter. daripada memperbarui parameter θ dengan gradien pada iterasi sekarang saja, metode *stochastic gradient descent* dengan momentum memperbarui parameter θ dengan gradien pada iterasi sekarang dan gradien pada iterasi sebelumnya. gradien yang digunakan untuk melakukan pembaruan parameter θ *exponential moving average* dari gradien pada iterasi sekarang dan gradien pada iterasi sebelumnya. Persamaan 2.44 menunjukkan algoritma *stochastic gradient descent* dengan momentum dan Gambar 2.15 mengilustrasikan pembaruan parameter θ dengan momentum.

$$\theta^{(t+1)} = \theta^{(t)} - \eta m^{(t+1)}, \quad (2.44)$$

$$m^{(t+1)} = \beta_1 m^{(t)} + (1 - \beta_1) \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}). \quad (2.45)$$



Gambar 2.15: Ilustrasi dari pembaruan parameter θ dengan *stochastic gradient descent* dengan momentum.

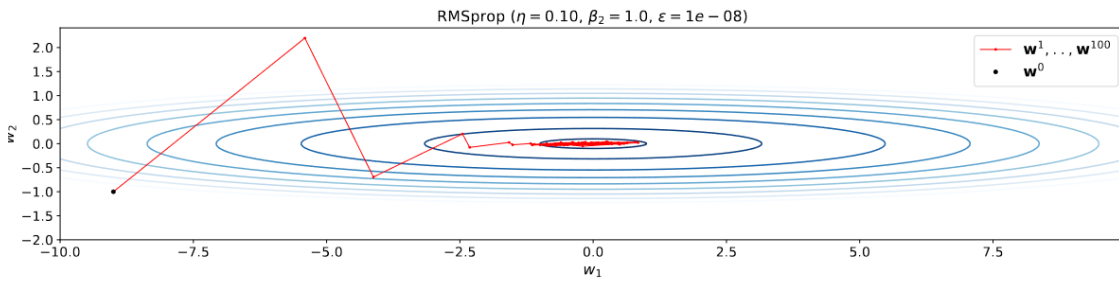
Sumber: (Geiger et al., 2022).

Metode lainnya yang dapat digunakan untuk mempercepat proses pembaruan parameter θ adalah metode *adaptive learning rate*. Metode *adaptive learning rate* mengubah

learning rate pada setiap parameter θ dengan membagi *learning rate* awal dengan *moving average* dari kuadrat gradien – biasanya disebut sebagai *running variance* – pada parameter θ tersebut. Pembagian antara gradien dan *running variance* tersebut dilakukan secara *element-wise*. Persamaan 2.46 menunjukkan algoritma *stochastic gradient descent* dengan *adaptive learning rate* dan Gambar 2.16 mengilustrasikan pembaruan parameter θ dengan *adaptive learning rate*.

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)})}{\sqrt{v^{(t+1)}} + \epsilon}, \quad (2.46)$$

$$v^{(t+1)} = \beta_2 v^{(t)} + (1 - \beta_2) \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}) \odot \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}). \quad (2.47)$$

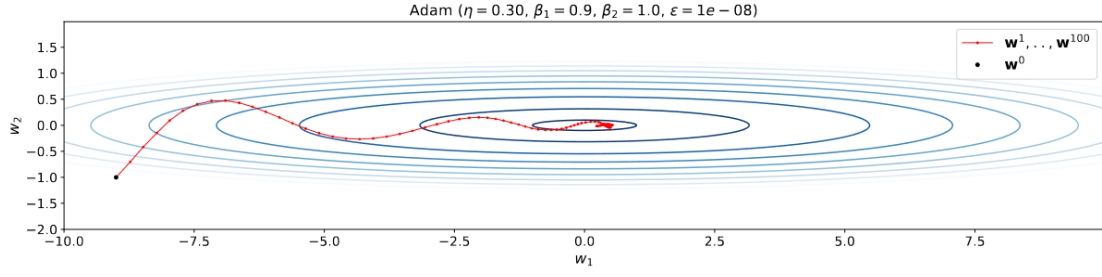


Gambar 2.16: Ilustrasi dari pembaruan parameter θ dengan *stochastic gradient descent* dengan *adaptive learning rate*.

Sumber: (Geiger et al., 2022).

faktor ϵ yang ditambahkan pada Persamaan 2.46 digunakan untuk menghindari pembagian dengan nol pada awal iterasi karena inisialisasi awal vektor $v^{(0)}$ adalah nol.

Terakhir, metode optimasi *Adaptive Moment Estimation* (Adam) menggabungkan metode *stochastic gradient descent* dengan momentum dan *adaptive learning rate*. Persamaan 2.48 menunjukkan algoritma *stochastic gradient descent* dengan Adam dan Gambar 2.17 mengilustrasikan pembaruan parameter θ dengan Adam. Persamaan 2.48 menunjukkan persamaan dari metode optimasi Adam dan Gambar 2.17 mengilustrasikan pembaruan parameter θ dengan Adam.



Gambar 2.17: Ilustrasi dari pembaruan parameter θ dengan Adam.
Sumber: (Geiger et al., 2022).

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta \hat{\mathbf{m}}^{(t+1)}}{\sqrt{\hat{\mathbf{v}}^{(t+1)} + \epsilon}}, \quad (2.48)$$

$$\hat{\mathbf{m}}^{(t+1)} = \frac{\mathbf{m}^{(t+1)}}{1 - \beta_1}, \quad (2.49)$$

$$\hat{\mathbf{v}}^{(t+1)} = \frac{\mathbf{v}^{(t+1)}}{1 - \beta_2}, \quad (2.50)$$

$$\mathbf{m}^{(t+1)} = \beta_1 \mathbf{m}^{(t)} + (1 - \beta_1) \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}), \quad (2.51)$$

$$\mathbf{v}^{(t+1)} = \beta_2 \mathbf{v}^{(t)} + (1 - \beta_2) \left(\nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}) \odot \nabla_{\theta} L_{\mathcal{B}}(\theta^{(t)}) \right). \quad (2.52)$$

Alasan dilakukan pembagian dengan $(1 - \beta_1)$ dan $(1 - \beta_2)$ Persamaan 2.49 dan Persamaan 2.50 adalah untuk menghilangkan bias pada *momentum* dan *running variance* pada awal iterasi.

2.4 Pembelajaran Representasi

2.4.1 Fungsi *Loss* pada Pembelajaran Representasi

Fungsi loss pada pembelajaran representasi biasanya disebut sebagai *ranking loss*. Meminimalkan fungsi *ranking loss* berarti memastikan bahwa *input-input* yang serupa berada lebih dekat daripada *input-input* yang tidak mirip. Sebagian besar fungsi loss pada pembelajaran representasi tidak memerlukan label kelas.

Salah satu fungsi loss yang digunakan pada penelitian ini adalah *N-pair loss* (van den Oord, Li, & Vinyals, 2018). *N-pair loss* dapat ditinjau sebagai klasifikasi multikelas dengan N-1 kelas berupa kelas negatif (kelas yang tidak mirip dengan *input* yang diberikan)

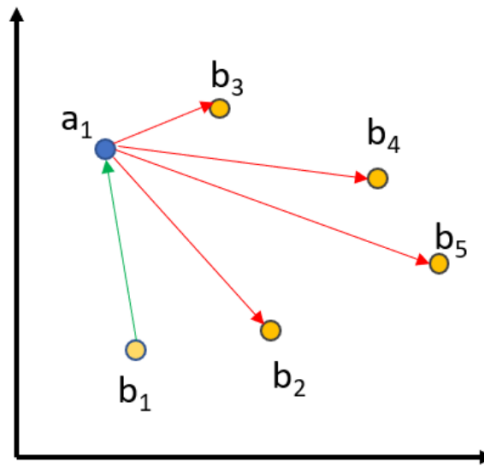
dan 1 kelas positif (kelas yang mirip dengan *input* yang diberikan). Untuk suatu *input* x dan *input* positif x^+ dan kumpulan *input* negatif $\{x_i^-\}_{i=1}^{N-1}$, N -pair loss dapat ditulis seperti pada Persamaan 2.53.

$$L(x, x^+, \{x_i^-\}_{i=1}^{N-1}) = -\log \frac{\exp(\text{sim}(f_{\text{model}}(x), f_{\text{model}}(x^+)))}{\sum_{i=1}^{N-1} \exp(\text{sim}(f_{\text{model}}(x), f_{\text{model}}(x_i^-))) + \exp(\text{sim}(f_{\text{model}}(x), f_{\text{model}}(x^+)))}, \quad (2.53)$$

dengan sim adalah fungsi yang mengukur keserupaan antara dua vektor (fungsi jarak). Pada penelitian ini, fungsi sim yang digunakan adalah fungsi *dot product* sehingga Persamaan 2.53 dapat ditulis ulang seperti pada Persamaan 2.54.

$$L(x, x^+, \{x_i^-\}_{i=1}^{N-1}) = -\log \frac{\exp(f_{\text{model}}(x)^\top f_{\text{model}}(x^+))}{\sum_{i=1}^{N-1} \exp(f_{\text{model}}(x)^\top f_{\text{model}}(x_i^-)) + \exp(f_{\text{model}}(x)^\top f_{\text{model}}(x^+))}. \quad (2.54)$$

Gambar 2.18 mengilustrasikan fungsi N -pair loss. Untuk pasangan teks yang relevan (a, b_1) , tujuannya adalah untuk meminimalkan jarak (fungsi sim) antara a dan b_1 sehingga jarak tersebut lebih kecil dibandingkan dengan jarak antara a dan b_i yang lain.



Gambar 2.18: Ilustrasi fungsi objektif N -pair loss. Untuk pasangan teks yang relevan (a, b_1) , tujuannya adalah untuk meminimalkan jarak antara a dan b_1 sehingga jarak tersebut lebih kecil dibandingkan dengan jarak antara a dan b_i yang lain.

Sumber: <https://www.sbert.net/>

BAB 3

BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER (BERT) UNTUK PEMERINGKATAN TEKS

3.1 Mekanisme *Attention*

Mekanisme *Attention* dapat ditinjau sebagai *Dictinoary Lookup*, yaitu untuk sebuah vektor kueri \mathbf{q} dan sekumpulan pasangan terurut vektor $\mathcal{KV} = \{(\mathbf{k}_1, \mathbf{v}_1), (\mathbf{k}_2, \mathbf{v}_2), \dots, (\mathbf{k}_n, \mathbf{v}_n)\}$, mekanisme *attention* akan mengembalikan vektor nilai \mathbf{v}_i yang memiliki vektor kunci \mathbf{k}_i yang cocok dengan vektor kueri \mathbf{q} . Persamaan 3.1 menunjukkan bagaimana mekanisme *attention* dilakukan.

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = [\alpha_1, \alpha_2, \dots, \alpha_n] \mathbf{V} \in \mathbb{R}^{d_v}, \quad (3.1)$$

dengan keterangan sebagai berikut:

$$\mathbf{K} = \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \vdots \\ \mathbf{k}_n \end{bmatrix} \in \mathbb{R}^{n \times d_k},$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{bmatrix} \in \mathbb{R}^{n \times d_v},$$

$$\alpha_i = \begin{cases} 1, & \text{jika } i = \arg \max_j f_{\text{attn}}(\mathbf{q}, \mathbf{k}_j) \\ 0, & \text{lainnya} \end{cases}.$$

$f_{\text{attn}}(\mathbf{q}, \mathbf{k})$ adalah fungsi atensi yang menghitung nilai atensi antara vektor kueri \mathbf{q} dan vektor kunci \mathbf{k} . α_i pada persamaan di atas disebut sebagai bobot atensi dan nilai $f_{\text{attn}}(\mathbf{q}, \mathbf{k})$ disebut sebagai nilai atensi.

Sebagai contoh, untuk $\mathbf{q} = [1, 2]$, $\mathcal{KV} = \{([2, 1], [1, 0]), ([1, 2], [0, 1])\}$ serta fungsi

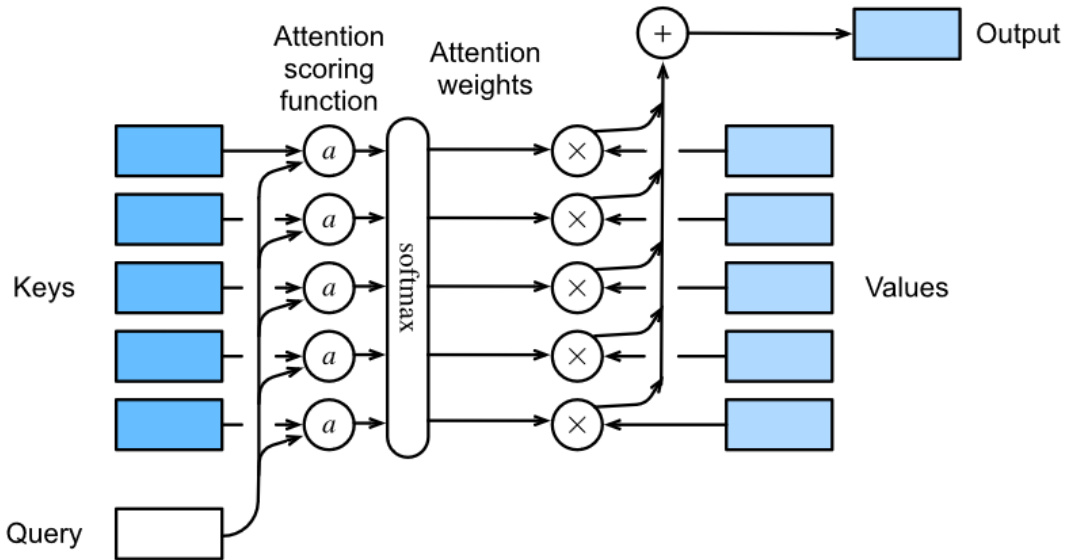
$f_{\text{attn}}(\mathbf{q}, \mathbf{k}) = \mathbf{q} \cdot \mathbf{k}$, nilai dari $\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V})$ adalah $[0, 1]$, karena nilai maksimal f_{attn} terjadi ketika $\mathbf{k} = [1, 2]$.

Mekanisme *attention* pada Persamaan 3.1 disebut sebagai *hard attention* karena hanya satu vektor nilai \mathbf{v}_i yang dipilih dari sekumpulan vektor nilai \mathbf{V} . Berbeda dengan *hard attention*, *soft attention* mengambil seluruh vektor nilai \mathbf{V} dan menghitung bobot α_i untuk setiap vektor nilai \mathbf{v}_i dengan fungsi *softmax*. Hasil dari *soft attention* adalah rata-rata terbobot dari seluruh vektor nilai \mathbf{V} . Persamaan 3.2 menunjukkan bagaimana mekanisme *soft attention* dilakukan.

$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = [\alpha_1, \alpha_2, \dots, \alpha_n] \mathbf{V} \in \mathbb{R}^{d_v}, \quad (3.2)$$

dengan nilai α_i yang dihitung seperti persamaan berikut:

$$\alpha_i(\mathbf{q}, \mathbf{k}_i) = \text{Softmax}_i(\alpha) = \frac{\exp(f_{\text{attn}}(\mathbf{q}, \mathbf{k}_i))}{\sum_{j=1}^n \exp(f_{\text{attn}}(\mathbf{q}, \mathbf{k}_j))}.$$



Gambar 3.1: Ilustrasi dari mekanisme *soft attention*.

Sumber: (A. Zhang et al., 2023).

Dengan rata-rata terbobot dari \mathbf{V} , transformasi *soft attention* terturunkan. Hal ini merupakan syarat *fundamental* yang harus dimiliki oleh sebuah model *deep learning*.

Sebagai contoh, hasil dari $\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V})$ untuk $\mathbf{q} = [1, 2]$, $\mathcal{KV} = \{([2, 1], [0, 1]), ([1, 2], [1, 0])\}$ serta fungsi $f_{\text{attn}}(\mathbf{q}, \mathbf{k}) = \mathbf{q} \cdot \mathbf{k}$ adalah $0.268[0, 1] +$

$0.732[1,0] = [0.732, 0.268]$ dengan $\alpha_1 = \frac{\exp(4)}{\exp(4)+\exp(5)} \approx 0.268$ dan $\alpha_2 = \frac{\exp(5)}{\exp(4)+\exp(5)} \approx 0.732$.

Pada kasus kumpulan kueri $Q = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m\}$, mekanisme *attention* untuk setiap triplet $(\mathbf{q}_i, \mathbf{K}, \mathbf{V})$ dapat dihitung secara bersamaan seperti yang ditunjukkan pada Persamaan 3.3.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{AV} \in \mathbb{R}^{m \times d_v}, \quad (3.3)$$

dengan:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_m \end{bmatrix} \in \mathbb{R}^{m \times d_k},$$

$$\mathbf{A} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

$$\alpha_{ij}(\mathbf{q}_i, \mathbf{k}_j) = \text{Softmax}_j(\alpha_i) = \frac{\exp(f_{\text{attn}}(\mathbf{q}_i, \mathbf{k}_j))}{\sum_{k=1}^n \exp(f_{\text{attn}}(\mathbf{q}_i, \mathbf{k}_k))} \in \mathbb{R},$$

dan α_{ij} adalah bobot yang menunjukkan bobot atensi antara vektor kueri \mathbf{q}_i dengan vektor kunci \mathbf{k}_j .

3.1.1 Attention Parametrik

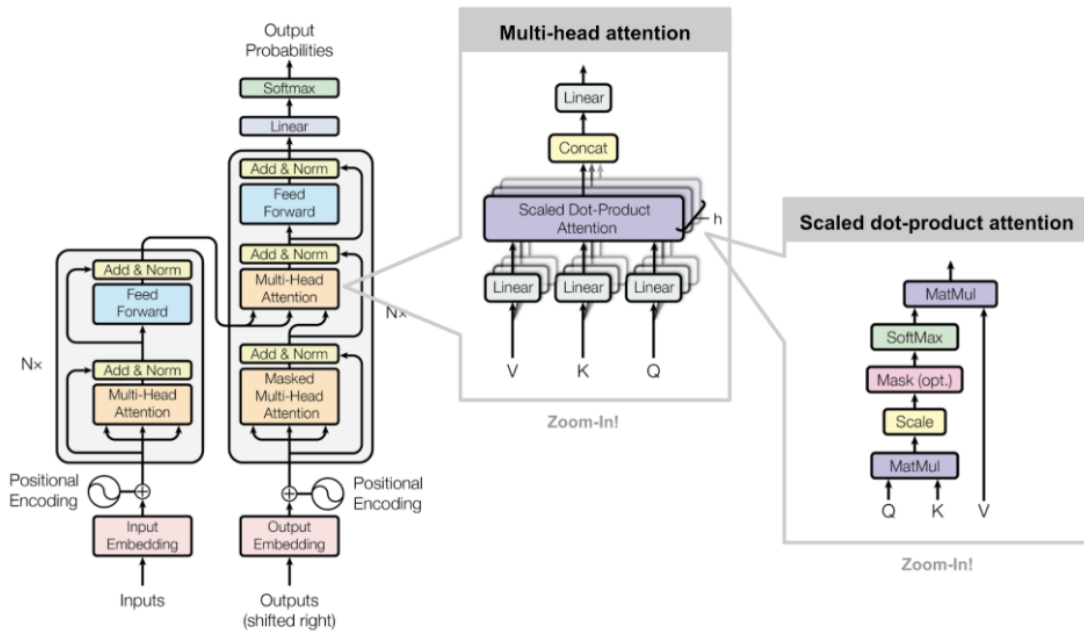
Mekanisme *attention* yang dilakukan oleh Vaswani et al. (2017) merupakan mekanisme *attention* parametrik. Pada mekanisme *attention* parametrik, nilai f_{attn} antar vektor kueri \mathbf{q} dan \mathbf{v} dibandingkan pada ruang vektor yang akan dipelajari (*learned embedding space*) daripada ruang vektor aslinya. Sebagai contoh, untuk suatu kueri $\mathbf{q} \in \mathbb{R}^{d_q}$, dan vektor kunci $\mathbf{k} \in \mathbb{R}^{d_k}$, *additive attention* yang diperkenalkan oleh Bahdanau, Cho, dan Bengio (2016) menghitung nilai keserupaan antara \mathbf{q} dan \mathbf{k} seperti pada Persamaan 3.4

$$f_{\text{attn}}(\mathbf{q}\mathbf{W}^q, \mathbf{k}\mathbf{W}^k) = (\mathbf{q}\mathbf{W}^q + \mathbf{k}\mathbf{W}^k)\mathbf{W}^{\text{out}} \in \mathbb{R}, \quad (3.4)$$

dengan $\mathbf{W}^q \in \mathbb{R}^{d_q \times d_{\text{attn}}}$, $\mathbf{W}^k \in \mathbb{R}^{d_k \times d_{\text{attn}}}$, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{attn}} \times 1}$ adalah matriks parameter yang akan diestimasi selama proses pelatihan. Contoh *attention* parametrik yang lebih sederhana adalah *dot-product attention*. Fungsi f_{attn} yang digunakan adalah perkalian titik antara \mathbf{q} dan \mathbf{k} di ruang vektor yang dipelajari (*learned embedding space*). Persamaan 3.5 menunjukkan bagaimana *dot-product attention* dihitung.

$$f_{\text{attn}}(\mathbf{q}\mathbf{W}^q, \mathbf{k}\mathbf{W}^k) = (\mathbf{q}\mathbf{W}^q)(\mathbf{k}\mathbf{W}^k)^\top. \quad (3.5)$$

3.2 Transformer



Gambar 3.2: Arsitektur *transformer* (Weng, 2018).

Transformer merupakan Arsitektur *deep learning* yang pertama kali diperkenalkan oleh Vaswani et al. (2017). Awalnya *Transformer* merupakan model *sequence to sequence* yang diperuntukkan untuk permasalahan mesin translasi neural (*neural machine translation*). Namun, sekarang *transformer* juga digunakan untuk permasalahan pemrosesan bahasa alami lainnya. model-model yang berarsitektur *transformer* menjadi model *state-of-the-art* untuk permasalahan pemrosesan bahasa alami lainnya, seperti *question answering*, *sentiment analysis*, dan *named entity recognition*.

Berbeda dengan arsitektur mesin translasi terdahulu, *transformer* tidak menggunakan *recurrent neural network* (RNN) atau *convolutional neural network* (CNN), melainkan

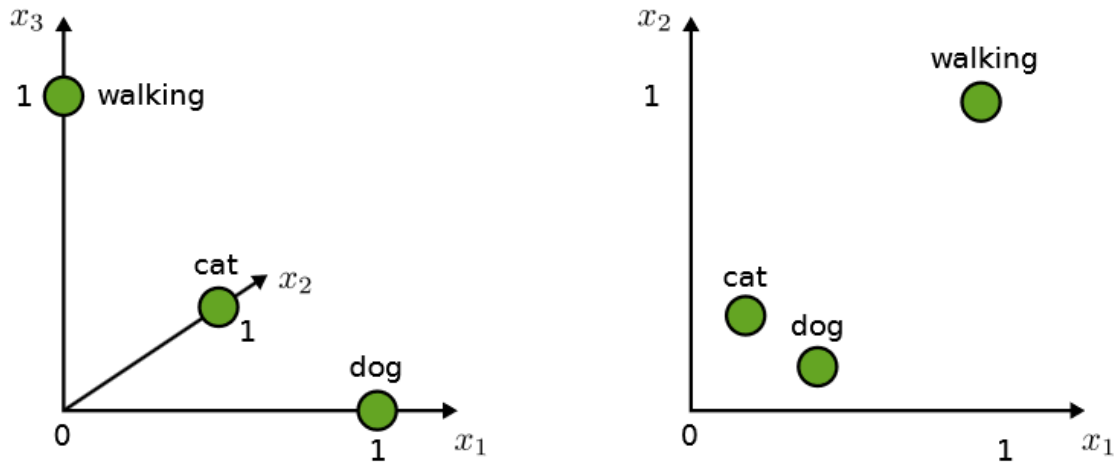
transformer adalah model *feed forward network* yang dapat memproses seluruh *input* pada barisan secara paralel. Untuk menggantikan kemampuan RNN dalam mempelajari ketergantungan antar *input* yang berurutan dan kemampuan CNN dalam mempelajari fitur lokal, transformer bergantung pada mekanisme *attention*.

Terdapat tiga jenis *attention* yang digunakan dalam model *transformer* (Vaswani et al., 2017):

1. *Encoder self-attention*: menggunakan barisan *input* yang berupa barisan token atau kata sebagai masukan untuk menghasilkan barisan representasi kontekstual, berupa vektor, dari *input*. Setiap representasi token tersebut memiliki ketergantungan dengan token lainnya dari barisan *input*.
2. *Decoder self-attention*: menggunakan barisan *target* yang berupa kalimat terjemahan parsial, barisan token, sebagai masukan untuk menghasilkan barisan representasi kontekstual (vektor) dari *target*. Setiap representasi token tersebut memiliki ketergantungan dengan token sebelumnya dalam urutan masukan.
3. *Decoder-encoder attention*: menggunakan barisan representasi kontekstual dari *input*, dan barisan representasi kontekstual dari *target* untuk menghasilkan token berikutnya yang merupakan hasil prediksi dari model. Barisan *target* yang digabung dengan token hasil prediksi tersebut akan menjadi barisan *target* untuk prediksi selanjutnya.

Arsitektur dari *transformer* terdiri dari pasangan encoder-decoder. Arsitektur dari *transformer* dapat dilihat pada Gambar 3.2. Lapisan *encoder* berfungsi untuk memahami konteks suatu kata dalam teks atau kalimat, sementara lapisan *decoder* digunakan untuk menyelesaikan masalah translasi menuju bahasa berbeda. Pada permasalahan klasifikasi seperti analisis sentimen dan pemeringkatan teks, lapisan *decoder* tidak digunakan. pada permasalahan tersebut, *output* dari lapisan *encoder* yang digunakan sebagai masukan untuk lapisan *classifier*. Subbab 3.2.1 hingga Subbab 3.2.8 menjelaskan arsitektur model *transformer encoder* dan berbagai mekanisme yang menyusun model *transformer*.

3.2.1 Token Embedding (Input Embedding)



Gambar 3.3: Ilustrasi dari representasi token. Gambar kiri menunjukkan representasi token dengan *one-hot encoding*, sedangkan gambar kanan menunjukkan representasi token dengan *token embedding*.

Sumber: (Geiger et al., 2022).

Perlu diingat kembali bahwa *input* dari *Attention* (dan tentunya *transformer*) adalah barisan vektor. Jika *Attention* ingin dapat digunakan pada permasalahan bahasa, barisan kata atau subkata (selanjutnya disebut token) harus terlebih dahulu diubah menjadi barisan vektor.

Representasi vektor dari token yang paling sederhana adalah dengan *one-hot encoding*. Andaikan $\mathcal{T} = \{t_1, t_2, \dots, t_{|\mathcal{T}|}\}$ adalah semua kemungkinan token yang mungkin muncul dalam permasalahan bahasa yang ingin diselesaikan. Untuk sembarang barisan token $t = (t_{i_1}, t_{i_2}, \dots, t_{i_L})$, representasi vektor dari token t_{i_j} adalah vektor $\mathbf{oh}_{i_j} = [0, \dots, 0, 1, 0, \dots, 0] \in \mathbb{R}^{|\mathcal{T}|}$, dengan nilai 1 pada indeks ke j dan nilai 0 pada indeks lainnya. *One-hot encoding* tentunya memiliki kelemahan:

1. Vektor yang dihasilkan adalah *sparse vector*, dan ukuran vektor yang dihasilkan cukup besar, yaitu $|\mathcal{T}|$.
2. Representasi token yang buruk. Operasi vektor yang dilakukan pada *one-hot encoding* tidaklah bermakna. Misalnya, Jarak antar token akan selalu sama pada *one-hot encoding*, yaitu $\sqrt{2}$.

Untuk Mengatasi kekurangan dari representasi *one-hot encoding*, representasi yang digunakan adalah vektor padat yang akan dipelajari ketika proses pelatihan. Misalkan

$\mathbf{E}_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}| \times d_{\text{token}}}$ adalah matriks parameter yang merupakan representasi vektor padat dari seluruh token ada. Persamaan 3.6 hingga Persamaan 3.7 menunjukkan bagaimana representasi vektor dari barisan suatu token t dihitung.

$$\text{Embed}(t) = \mathbf{E}_t = \begin{bmatrix} \mathbf{e}_{i_1} \\ \mathbf{e}_{i_2} \\ \vdots \\ \mathbf{e}_{i_L} \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}}, \quad (3.6)$$

$$\mathbf{e}_{i_j} = \mathbf{oh}_{i_j} \mathbf{E}_{\mathcal{T}} \in \mathbb{R}^{d_{\text{token}}}. \quad (3.7)$$

Gambar 3.3 mengilustrasikan perbedaan antara *one-hot encoding* dan *token embedding*. Pada representasi token dengan vektor padat, vektor yang secara semantik atau sintaksis mirip akan memiliki jarak yang lebih dekat. Selain itu, biasanya representasi token dengan vektor padat memiliki dimensi d_{token} yang lebih kecil daripada *one-hot encoding* yang memiliki dimensi $|\mathcal{T}|$.

3.2.2 Scaled Dot-Product Attention

Scaled dot-product attention adalah mekanisme *Attention* parametrik yang digunakan dalam *transformers*. *Scaled dot-product attention* menghitung keserupaan antara vektor kueri \mathbf{q} dan vektor kunci \mathbf{k} pada ruang vektor yang dipelajari (*learned embedding space*) dengan fungsi keserupaan $f_{\text{attn}}(\mathbf{q}\mathbf{W}^q, \mathbf{k}\mathbf{W}^k)$ adalah perkalian titik antara $\mathbf{q}\mathbf{W}^q$ dan $\mathbf{k}\mathbf{W}^k$ yang kemudian dibagi dengan $\sqrt{d_{\text{attn}}}$, seperti yang ditunjukkan Persamaan 3.8.

$$f_{\text{attn}}(\mathbf{q}\mathbf{W}^q, \mathbf{k}\mathbf{W}^k) = \frac{\mathbf{q}\mathbf{W}^q(\mathbf{k}\mathbf{W}^k)^\top}{\sqrt{d_{\text{attn}}}} \in \mathbb{R}. \quad (3.8)$$

dengan $\mathbf{W}^q \in \mathbb{R}^{d_q \times d_{\text{attn}}}$, $\mathbf{W}^k \in \mathbb{R}^{d_k \times d_{\text{attn}}}$ adalah matriks parameter dan d_{attn} adalah dimensi dari *learned embedding space* yang digunakan untuk perhitungan nilai atensi.

Pembagian dengan $\sqrt{d_{\text{attn}}}$ dilakukan untuk menjaga variansi dari nilai atensi $\mathbf{q}\mathbf{W}^q(\mathbf{k}\mathbf{W}^k)^\top$ tetap serupa dengan variansi $\mathbf{q}\mathbf{W}^q$ dan $\mathbf{k}\mathbf{W}^k$. Tanpa pembagian $\sqrt{d_{\text{attn}}}$, variansi dari nilai atensi akan memiliki faktor tambahan $\sigma^2 d_{\text{attn}}$, seperti yang ditunjukkan

pada Persamaan 3.9 hingga Persamaan 3.10.

$$\mathbf{qW}^q \sim \mathcal{N}(0, \sigma^2) \text{ dan } \mathbf{kW}^k \sim \mathcal{N}(0, \sigma^2), \quad (3.9)$$

$$\text{Var}(\mathbf{qW}^q(\mathbf{kW}^k)^\top) = \sum_{i=1}^{d_{\text{attn}}} \text{Var}\left((\mathbf{qW}^q)_i((\mathbf{kW}^k)_i)^\top\right) = \sigma^4 d_{\text{attn}}. \quad (3.10)$$

Akibatnya, untuk nilai d_{attn} yang cukup besar, akan terdapat satu elemen atensi acak $(\mathbf{qW}^q(\mathbf{kW}^k)^\top)_i$ sehingga $|(\mathbf{qW}^q(\mathbf{kW}^k)^\top)_i| \gg |(\mathbf{qW}^q(\mathbf{kW}^k)^\top)_j|$ untuk sembarang nilai atensi lainnya. Jika faktor d_{attn} tidak dihilangkan, *softmax* dari nilai atensi akan jenuh ke 1 untuk satu elemen acak tersebut dan 0 untuk elemen lainnya – atau sebaliknya. Akibatnya, gradien pada fungsi *softmax* akan mendekati nol sehingga model tidak dapat belajar parameter dengan baik.

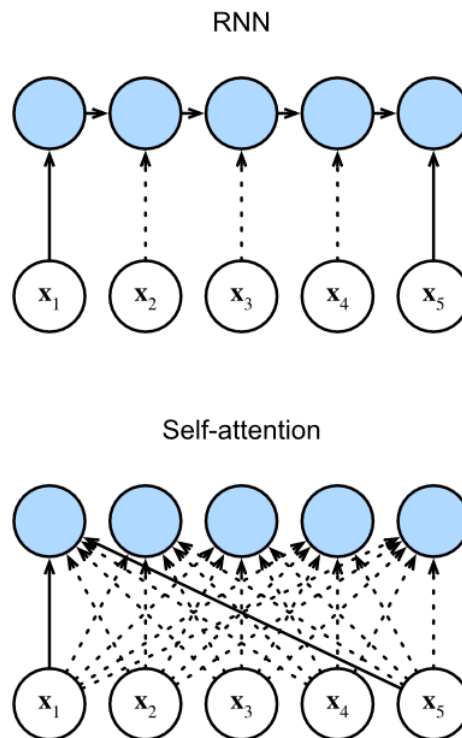
Dengan *scaled dot product attention*, tidak ada faktor d_{attn} pada variansi nilai atensi. faktor σ^4 pada Persamaan 3.11 tidak menjadi masalah karena dengan *normalization layer* yang dijelaskan pada Subbab 3.2.7 mengakibatkan $\sigma^2 \approx 1$ sehingga $\sigma^4 \approx \sigma^2 \approx 1$.

$$(\text{scaled dot product attention}) \text{Var}\left(\frac{\mathbf{qW}^q(\mathbf{kW}^k)^\top}{\sqrt{d_{\text{attn}}}}\right) = \frac{\sigma^4 d_{\text{attn}}}{d_{\text{attn}}} = \sigma^4 \quad (3.11)$$

Terakhir, untuk kumpulan vektor kueri $Q = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m\}$, dan kumpulan vektor kunci dan nilai $\mathcal{KV} = \{(\mathbf{k}_1, \mathbf{v}_1), (\mathbf{k}_2, \mathbf{v}_2), \dots, (\mathbf{k}_n, \mathbf{v}_n)\}$, *scaled dot product attention* dapat dihitung secara bersamaan seperti pada Persamaan 3.12.

$$\text{Attention}(\mathbf{QW}^q, \mathbf{KW}^k, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{QW}^q(\mathbf{KW}^k)^\top}{\sqrt{d_{\text{attn}}}}\right)\mathbf{V} \in \mathbb{R}^{m \times d_v}. \quad (3.12)$$

3.2.3 Self-Attention



Gambar 3.4: Perbandingan RNN dan *self-attention* dalam menghasilkan representasi vektor kontekstual. Pada RNN, representasi vektor kontekstual setiap token bergantung pada perhitungan token sebelumnya. Pada *self-attention*, representasi vektor kontekstual setiap token dihitung secara independen dan paralel.
Sumber: (A. Zhang et al., 2023).

Self-Attention layer adalah *layer* yang digunakan *transformer* untuk menghasilkan representasi vektor yang kontekstual dari barisan token *input*. Berbeda dengan RNN dalam menghasilkan representasi vektor kontekstual, *self-attention* tidak memerlukan ketergantungan sekuensial, yang berarti representasi vektor kontekstual setiap tokennya dapat dihitung secara independen dan paralel. Gambar 3.4 menggambarkan perbedaan kedua arsitektur dalam menghasilkan representasi vektor kontekstual. Kemampuan Paralelisme dari *self-attention* membuat proses komputasi menjadi lebih cepat pada *hardware* yang mendukung paralelisme.

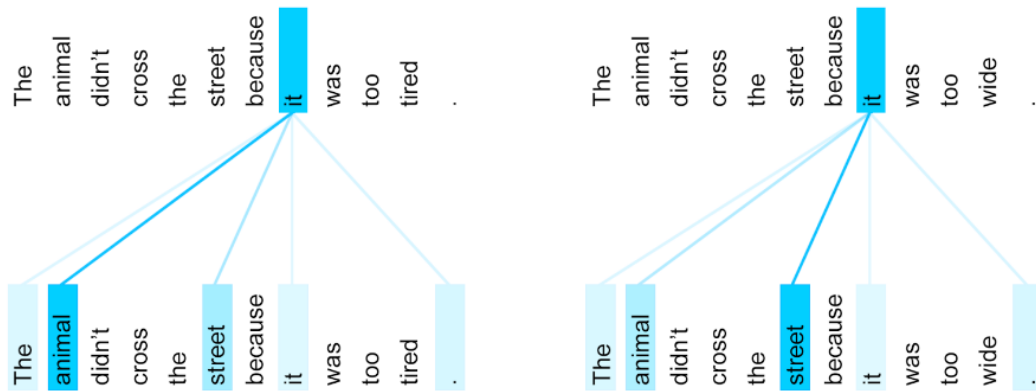
Perhitungan *self-attention* pada *transformer* yang digunakan adalah *scaled dot product attention* yang telah dijelaskan pada Subbab 3.2.2. Pada *self-attention*, kumpulan vektor kueri \mathbf{Q} , vektor kunci \mathbf{K} , dan vektor nilai \mathbf{V} adalah vektor yang sama, yaitu *embedding* dari token \mathbf{E} yang dijelaskan pada Subbab 3.2.1. Selain itu, Dimensi dari *learned embedding space* d_{attn} yang digunakan untuk perhitungan nilai atensi adalah d_{token} yaitu

dimensi dari *token embedding*. Persamaan 3.13 menunjukkan bagaimana *self-attention* dihitung.

$$\begin{aligned}\text{Self-Attention}(\mathbf{E}) &= \text{Attention}(\mathbf{E}\mathbf{W}^q, \mathbf{E}\mathbf{W}^k, \mathbf{E}\mathbf{W}^v) \\ &= \text{Softmax}\left(\frac{\mathbf{E}\mathbf{W}^q(\mathbf{E}\mathbf{W}^k)^\top}{\sqrt{d_{\text{attn}}}}\right)(\mathbf{E}\mathbf{W}^v) \in \mathbb{R}^{L \times d_{\text{token}}},\end{aligned}\quad (3.13)$$

dengan $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v \in \mathbb{R}^{d_{\text{token}} \times d_{\text{token}}}$ adalah matriks bobot.

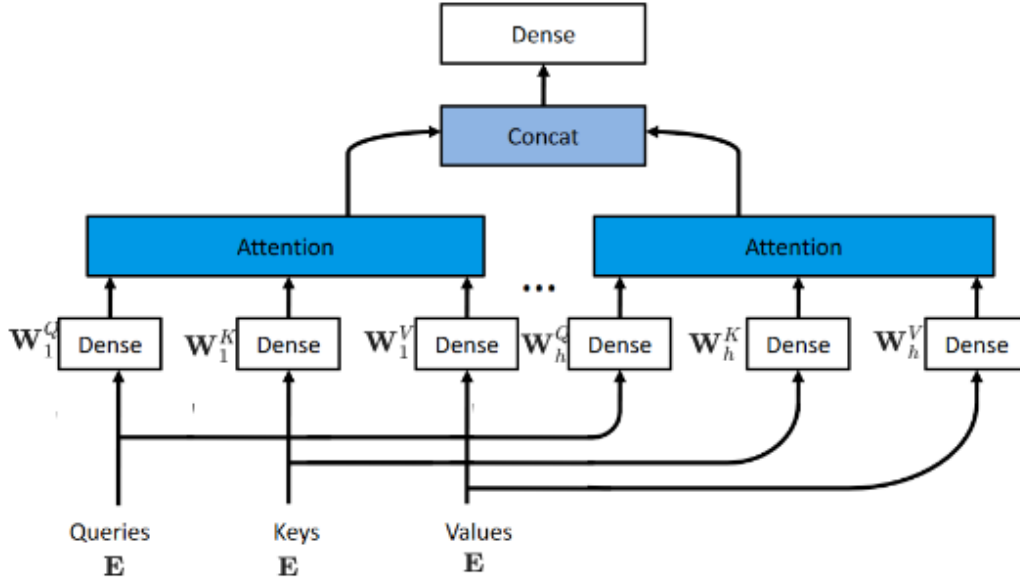
self-attention dapat dikonsepsikan sebagai proses pembentukan representasi token yang kontekstual. Untuk setiap tokennya, *self-attention* menghitung keserupaan antara token $\mathbf{E}\mathbf{W}^q$ dengan seluruh token lainnya $\mathbf{E}\mathbf{W}^k$ dengan *scaled dot product attention*. Hasil dari *scaled dot product attention* adalah vektor yang menunjukkan bobot atensi dari token tersebut terhadap token lainnya. Bobot atensi tersebut kemudian digunakan untuk menghitung rata-rata terbobot dari seluruh token lainnya ($\mathbf{E}\mathbf{W}^v$). Hasil dari rata-rata terbobot tersebut adalah representasi vektor kontekstual dari token tersebut. Gambar 3.5 adalah contoh dari *self-attention* yang menghasilkan representasi vektor kontekstual pada token *it*. Pada Gambar 3.5 kiri token *it* memiliki bobot atensi yang tinggi terhadap token *animal* dan *street* sehingga representasi vektor kontekstual dari token *it* akan memiliki nilai yang serupa dengan representasi token *animal*. Di lain sisi, token *it* pada Gambar 3.5 memiliki bobot atensi yang tinggi terhadap token *street*.



Gambar 3.5: Ilustrasi *self-attention* dalam menghasilkan representasi vektor kontekstual dari barisan token. Representasi vektor dari token *it* akan bergantung terhadap barisan token *input*.

Sumber: (Murphy, 2022)

3.2.4 Multi-Head Self-Attention



Gambar 3.6: Ilustrasi *multi-head self-attention* pada *transformer*. *Multi-head self-attention* menghitung *self-attention* sebanyak h kali pada subruang yang berbeda.

Sumber: (Murphy, 2022), telah diolah kembali.

Multi-Head Self-Attention adalah arsitektur pada *transformer* untuk melakukan mekanisme *self-attention* beberapa kali pada subruang (*learned embedded space*) yang berbeda. dengan melakukan hal tersebut, diharapkan bahwa model dapat menangkap relasi atau keserupaan antar token dari sudut pandang yang berbeda.

Secara teknis, *embedding* dari barisan token \mathbf{E} akan dipetakan sebanyak h kali dengan *linear layer* yang kemudian hasil *attention* dari setiap *head* akan digabungkan dan dilakukan transformasi sekali lagi dengan *linear layer*. Persamaan 3.14 menunjukkan bagaimana *multi-head self-attention* dihitung.

$$\text{MHSA}(\mathbf{E}) = [\text{head}_1 | \dots | \text{head}_h] \mathbf{W}^O \in \mathbb{R}^{L \times d_{\text{token}}}, \quad (3.14)$$

$$\text{head}_i = \text{Self-Attention}_i(\mathbf{E}) = \text{Softmax}\left(\frac{\mathbf{E}\mathbf{W}_i^q (\mathbf{E}\mathbf{W}_i^k)^\top}{\sqrt{d_{\text{token}}/h}}\right) \mathbf{E}\mathbf{W}_i^v \in \mathbb{R}^{L \times \frac{d_{\text{token}}}{h}}, \quad (3.15)$$

dengan $\mathbf{W}_i^q, \mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^{\frac{d_{\text{token}}}{h} \times \frac{d_{\text{token}}}{h}}$, $\mathbf{W}^O \in \mathbb{R}^{d_{\text{token}} \times d_{\text{token}}}$ adalah matriks bobotnya.

Perhatikan bahwa dimensi dari *learned embedding space* menjadi $\frac{d_{\text{token}}}{h}$ untuk setiap *head*-nya. Hal ini dilakukan untuk menjaga dimensi dari *output* terakhir tetap sama de-

ngan dimensi dari *input*, yaitu d_{token} . Selain itu, justifikasi lainnya yang dapat dibuat adalah setiap *head* hanya perlu menggunakan dimensi yang lebih kecil dari d_{token} untuk menangkap ketergantungan antar-token (pi tau, 2023).

3.2.5 Positional Encoding

Mekanisme *self-attention* yang dijelaskan sebelumnya tidak memperhatikan informasi mengenai urutan token selama proses komputasinya. Representasi vektor kontekstual dari suatu token akan sama meskipun urutan tokennya berbeda. Lebih tepatnya, mekanisme *self-attention* bersifat *permutation equivariant*, yaitu untuk *token embedding* \mathbf{E} dan matriks permutasi \mathbf{P}_π , Persamaan 3.16 terpenuhi.

$$\text{Self-Attention}(\mathbf{E}\mathbf{P}_\pi) = \text{Self-Attention}(\mathbf{E})\mathbf{P}_\pi \quad (3.16)$$

Namun, urutan dari token penting dalam pemrosesan bahasa alami. Kalimat `saya makan nasi` dan `nasi makan saya` memiliki makna yang berbeda. Oleh karena itu, informasi mengenai urutan token haruslah diperhatikan dalam pemrosesan bahasa alami.

Vaswani et al. (2017) menambahkan informasi posisi dengan Menjumlahkan *token embedding* \mathbf{E} dengan suatu matriks *positional encoding* \mathbf{PE} . setiap entri dari \mathbf{PE} adalah fungsi sinusoidal dari posisi token dan dimensi dari *token embedding* seperti yang ditunjukkan pada Persamaan 3.17.

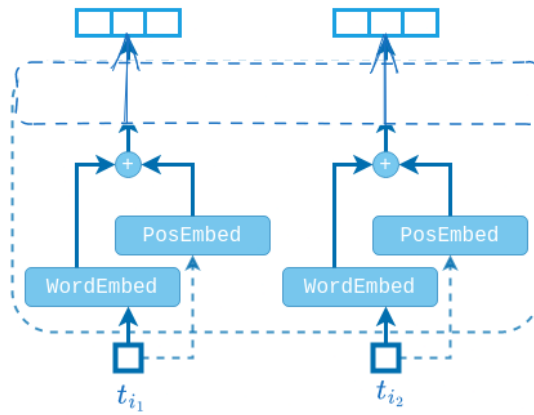
$$\text{PE}_{\text{pos},i} = \begin{cases} \sin\left(\frac{\text{pos}}{10000^{i/d_{\text{token}}}}\right) & \text{jika } i \bmod 2 = 0, \\ \cos\left(\frac{\text{pos}}{10000^{(i-1)/d_{\text{token}}}}\right) & \text{lainnya.} \end{cases} \quad (3.17)$$

berbeda dengan Vaswani et al. (2017), Devlin, Chang, Lee, dan Toutanova (2018) menggunakan matriks parameter $\mathbf{W}^{pe} \in \mathbb{R}^{L_{\text{max}} \times d_{\text{token}}}$ untuk menghitung matriks *positional encoding* $\mathbf{PE} \in \mathbb{R}^{L \times d_{\text{token}}}$ seperti yang ditunjukkan pada Persamaan 3.18 hingga Persamaan 3.19. Kekurangan dari pendekatan ini adalah model tidak dapat melakukan inferensi pada barisan token yang lebih panjang dari L_{max} . Gambar 3.7 mengilustrasikan

positional encoding pada transformer.

$$\mathbf{pe}_i = [0, 0, \dots, \underbrace{1}_{\text{indeks ke-}i}, 0, \dots, 0] \mathbf{W}^{pe} \in \mathbb{R}^{d_{\text{token}}}, \quad (3.18)$$

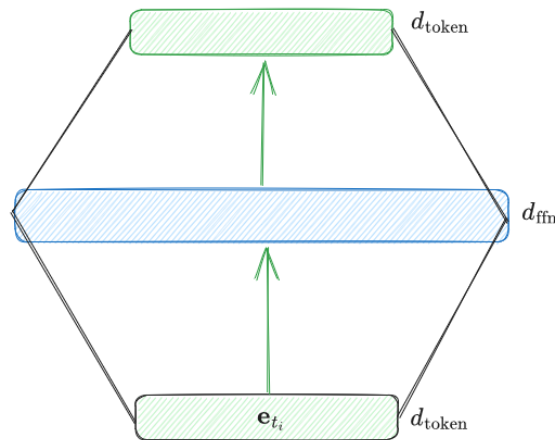
$$\text{pos}(t) = \mathbf{PE} = \begin{bmatrix} \mathbf{pe}_1 \\ \mathbf{pe}_2 \\ \vdots \\ \mathbf{pe}_L \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}}. \quad (3.19)$$



Gambar 3.7: Ilustrasi dari *positional encoding* pada *transformer*. *Positional encoding* ditambahkan pada *token embedding* sebelum dijadikan *input* untuk *transformer*.

Sumber: (pi tau, 2023), telah diolah kembali.

3.2.6 Position-wise Feed-Forward Network



Gambar 3.8: Ilustrasi *position-wise feed-forward network* pada *transformer*.

Position-wise Feed-Foward Network adalah *feed foward network* dengan dua kali transformasi linear dan sebuah fungsi aktivasi ReLU di antaranya. Gambar 3.8 menunjukkan ilustrasi dari *position-wise feed-forward network* dan Persamaan 3.20 menunjukkan Transformasi yang dilakukan oleh *position-wise feed-forward network*.

$$\text{FFN}(\mathbf{X}) = \max(0, \mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \in \mathbb{R}^{L \times d_{\text{token}}}, \quad (3.20)$$

dengan $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{token}} \times d_{\text{ffn}}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ffn}} \times d_{\text{token}}}$, $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{ffn}}}$, $\mathbf{b}_2 \in \mathbb{R}^{d_{\text{token}}}$ adalah matriks bobot dan bias. d_{ffn} adalah dimensi dari *feed forward network* yang digunakan. Vaswani et al. (2017) menggunakan $d_{\text{ffn}} = 2048$.

3.2.7 Koneksi Residu dan *Layer Normalization*

Pembaruan parameter model dilakukan pada semua *layer* secara serentak setiap iterasi *gradient descent*. Ketika parameter suatu *layer* mengalami pembaruan, distribusi dari *output* yang dihasilkan *layer* tersebut juga akan berubah pada iterasi selanjutnya. *Layer-layer* selanjutnya harus beradaptasi karena distribusi *input* dari *layer* tersebut berubah. Fenomena ini disebut *internal covariate shift* yang mengakibatkan proses pencarian parameter menjadi lebih lambat.

Layer Normalization berfungsi untuk mencegah masalah *internal covariate shift* di atas dengan membatasi distribusi nilai *output* – yang nantinya menjadi *input* pada *layer* selanjutnya – sehingga memiliki variansi 1 dan mean 0. Justifikasi lainnya di balik penggunaan *layer normalization* adalah variansi dari *input* untuk *self-attention layer* haruslah 1 (lihat Subbab 3.2.2), sehingga variansi dari bobot atensi $\text{Softmax}\left(\frac{\mathbf{E}\mathbf{W}^q(\mathbf{E}\mathbf{W}^k)^\top}{\sqrt{d_{\text{token}}}}\right)$ akan 1 juga. Persamaan 3.21 menunjukkan proses kerja dari *layer normalization*.

$$\begin{aligned} \text{LayerNorm}(\mathbf{X}) &= (\mathbf{X} - \boldsymbol{\mu}) \odot \frac{1}{\boldsymbol{\sigma}} \\ &= \begin{bmatrix} \frac{x_{11}-\mu_1}{\sigma_1} & \frac{x_{12}-\mu_1}{\sigma_1} & \dots & \frac{x_{1,d_{\text{token}}}-\mu_1}{\sigma_1} \\ \frac{x_{21}-\mu_2}{\sigma_2} & \frac{x_{22}-\mu_2}{\sigma_2} & \dots & \frac{x_{2,d_{\text{token}}}-\mu_2}{\sigma_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{L1}-\mu_L}{\sigma_L} & \frac{x_{L2}-\mu_L}{\sigma_L} & \dots & \frac{x_{L,d_{\text{token}}}-\mu_L}{\sigma_L} \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}}, \end{aligned} \quad (3.21)$$

dengan keterangan berikut:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1,d_{\text{token}}} \\ x_{21} & x_{22} & \dots & x_{2,d_{\text{token}}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & \dots & x_{L,d_{\text{token}}} \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}},$$

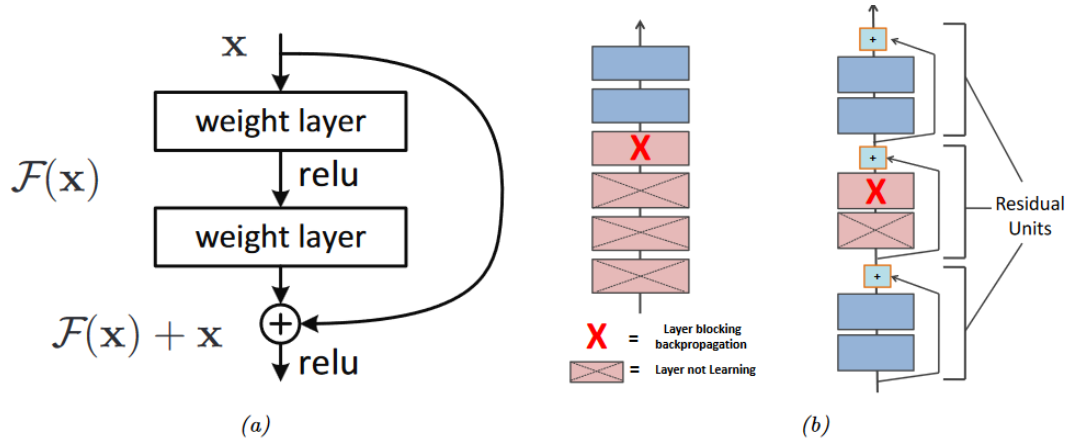
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \dots & \mu_1 \\ \vdots & \ddots & \vdots \\ \mu_L & \dots & \mu_L \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}},$$

$$\frac{1}{\boldsymbol{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_1} & \dots & \frac{1}{\sigma_1} \\ \vdots & \ddots & \vdots \\ \frac{1}{\sigma_L} & \dots & \frac{1}{\sigma_L} \end{bmatrix} \in \mathbb{R}^{L \times d_{\text{token}}},$$

$$\mu_i = \frac{1}{d_{\text{token}}} \sum_{j=1}^{d_{\text{token}}} x_{ij}, \quad i = 1, \dots, L,$$

$$\sigma_i = \sqrt{\frac{1}{d_{\text{token}}} \sum_{j=1}^{d_{\text{token}}} (x_{ij} - \mu_i)^2} \quad i = 1, \dots, L,$$

\odot = element-wise product.



Gambar 3.9: Ilustrasi koneksi residu.

Sumber: (Murphy, 2022)

Koneksi Residu adalah koneksi yang menghubungkan *output* dari suatu *layer* dengan *input* dari *layer* selanjutnya. Koneksi residu digunakan untuk mengatasi masalah *van-*

ishing gradient yang terjadi pada *deep neural network* dengan memperbaiki *flow gradient* dari model. Persamaan matematis dari koneksi residu dijelaskan seperti pada Persamaan 3.22.

$$f'_l(\mathbf{x}) = f_l(\mathbf{x}) + \mathbf{x}, \quad (3.22)$$

dengan $f_l(\mathbf{x})$ adalah suatu *layer* atau kumpulan *layer* pada *deep neural network*. Pada *transformer*, *residual connection* digunakan sebelum *layer normalization*.

3.2.8 Transformer Encoder

Dengan menggunakan *multi-head self-attention layer*, *position-wise feed-forward network layer*, dan *layer normalization* dan *residual connection* yang sudah dijelaskan sebelumnya, blok *encoder* pada *transformer encoder* dapat ditulis seperti pada Persamaan 3.26 hingga Persamaan 3.27.

$$\text{EncoderBlock}(\mathbf{X}) = \mathbf{Z}_2, \quad (3.23)$$

$$\mathbf{Z}_2 = \text{LayerNorm}(\overbrace{\text{FFN}(\mathbf{Z}_1) + \mathbf{Z}_1}^{\text{Koneksi Residu}}), \quad (3.24)$$

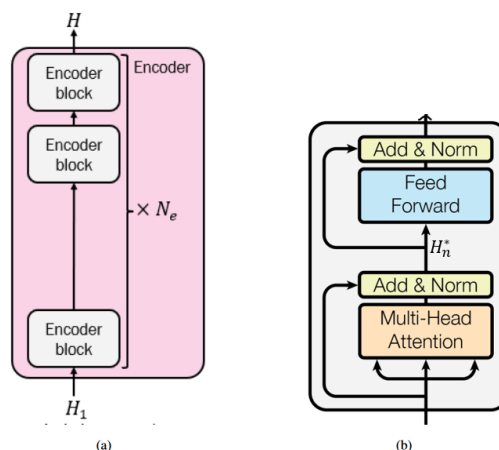
$$\mathbf{Z}_1 = \text{LayerNorm}(\overbrace{\text{MHSA}_h(\mathbf{X}), + \mathbf{X}}^{\text{Koneksi Residu}}), \quad (3.25)$$

dengan $\mathbf{X} \in \mathbb{R}^{L \times d_{\text{token}}}$ adalah *input* dari blok *transformer*.

Terakhir, *transformer encoder* adalah komposisi dari beberapa blok *encoder*. Untuk *input* token $t = (t_1, t_2, \dots, t_L)$, *transformer encoder* menghasilkan representasi vektor kontekstual dari setiap tokennya ditunjukkan pada Persamaan 3.26 hingga Persamaan 3.27.

$$\mathbf{E} = \text{embed}(t) + \text{pos}(t), \quad (3.26)$$

$$\text{Encoder}(\mathbf{E}) = \text{EncoderBlock}_n(\text{EncoderBlock}_{n-1}(\dots(\text{EncoderBlock}_1(\mathbf{E}))). \quad (3.27)$$



Gambar 3.10: Ilustrasi *transformer encoder*. (a) *Transformer encoder*, (b) *encoder blok*.

3.3 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah *transformers encoder* yang telah dilatih sebelumnya (*pre-trained*) pada tugas *Masked Language Model* dan *Next Sentence Prediction* (Devlin et al., 2018). *Pre-training* BERT dilakukan dengan menggunakan korpus teks yang besar dan dilakukan secara *self-supervised* untuk mempelajari informasi umum tentang statistik bahasa. penggunaan BERT mengikuti prinsip *transfer learning*, yaitu model yang sudah dilatih sebelumnya pada tugas tertentu – dengan jumlah data yang besar – dapat digunakan untuk tugas lainnya dengan hanya menggunakan sedikit data latih. Model BERT_{BASE} tersusun atas 12 blok *encoder* dengan dimensi *token embedding* sebesar 768, 12 *attention heads* pada setiap blok *encoder*-nya dan 110 juta parameter.

3.3.1 Representasi Input

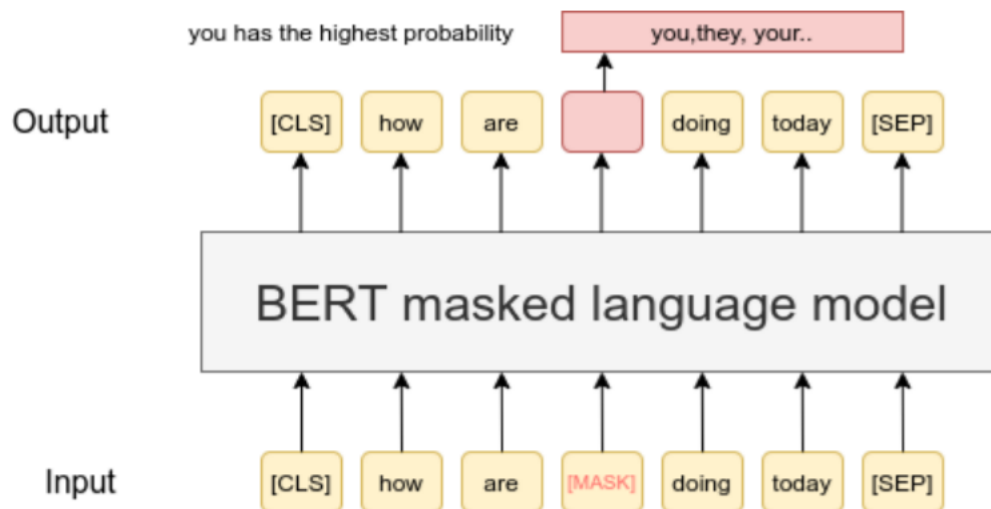
Representasi *input* dari model BERT tidak hanya terdiri dari *token embedding* dan *positional encoding* seperti pada *transformer encoder* yang dijelaskan pada Subbab 3.2.8

3.3.2 *pre-training* BERT

Pada tahap *pre-training*, BERT dilatih pada dua tugas *self-supervised* dengan jumlah data yang besar, yaitu *Masked Language Model* dan *Next Sentence Prediction* yang masing-masing akan dijelaskan pada Subbab 3.3.2.1 dan Subbab 3.3.2.2. proses *pre-training*

menggunakan paragraf-paragraf pada korpus teks yang besar, yaitu Wikipedia dengan 2.5 Miliar kata dan BookCorpus dengan 800 Juta kata.

3.3.2.1 *Masked Language Model (MLM)*



Gambar 3.11: Ilustrasi *Masked Language Modeling* (MLM) pada BERT. sebuah kata (token) secara acak di-hilangkan (*mask*) dan model diminta untuk menebak kata yang dihilangkan tersebut.

Tugas *Masked Language Model*(MLM) adalah tugas untuk memprediksi token yang dihilangkan (*masking*) pada suatu kalimat. Sebagai contoh pada kalimat Let's make [MASK] chicken! [SEP] It [MASK] great with orange sauce, model harus memprediksi token *fried* dan *tastes* pada token yang dihilangkan tersebut.

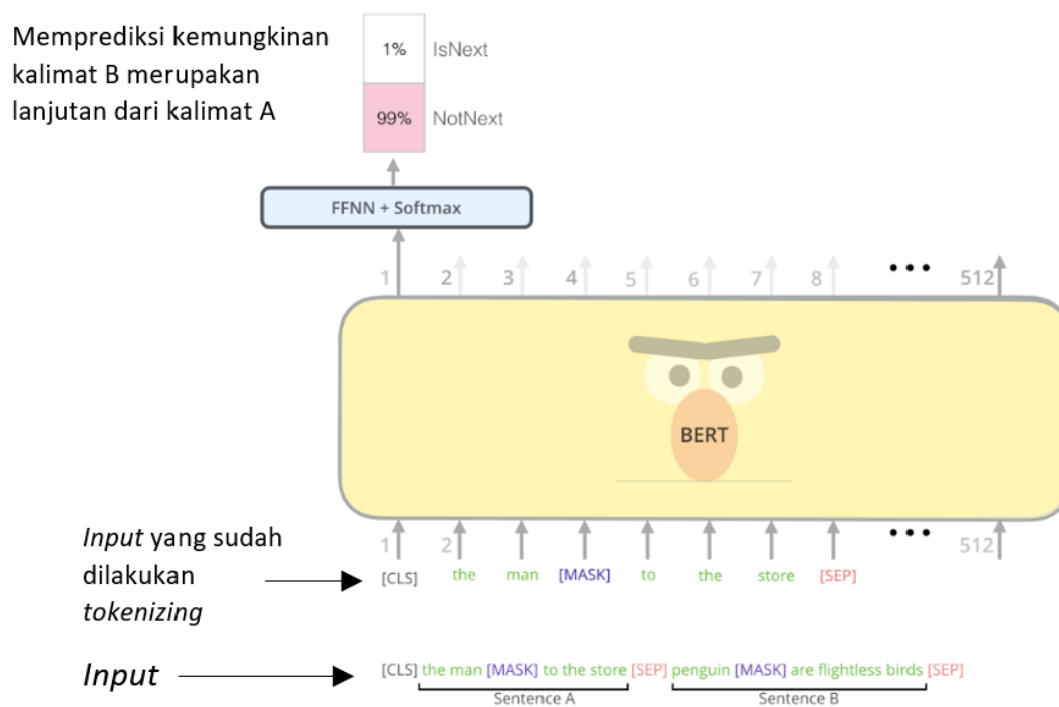
memprediksi kata yang dihilangkan memaksa *transformer* untuk memahami sintaks dan konteks dari kalimat tersebut. Sebagai contoh, model harus memahami bahwa kata sifat *red* sering terletak sebelum kata benda seperti *house* atau *car*, tetapi tidak sebelum kata kerja seperti *shout*. Selain itu, tugas ini membuat model untuk memperoleh pemahaman umum tentang bahasa alami. Misalnya, setelah dilatih, model akan memberikan probabilitas yang lebih tinggi untuk kata *train* yang hilang dalam kalimat [MASK] pulled into the station daripada kata *chicken*.

selama proses pelatihan MLM, sebanyak 15% dari semua token dipilih untuk dilakukan *masking*. 80% dari token yang terpilih akan diubah menjadi token [MASK], 10% menjadi token acak, dan 10% lainnya adalah token yang sama.

3.3.2.2 Next Sentence Prediction

pada tugas *Next Sentence Prediction*, model BERT dapat dilatih untuk memprediksi apakah kalimat kedua dari pasangan kalimat adalah kalimat berikutnya. dengan kata lain, untuk pasangan kalimat (q, d) , model BERT harus memprediksi apakah kalimat d adalah kalimat berikutnya dari kalimat q . *Input* dari tugas ini adalah $[CLS] \ t_{q_1}, t_{q_2}, \dots, t_{q_n} \ [SEP] \ t_{d_1}, t_{d_2}, \dots, t_{d_m} \ [SEP]$ dengan t_{q_i} adalah token dari kalimat q dan t_{d_i} adalah token dari kalimat d dan *output* dari tugas ini adalah variabel biner yang menunjukkan apakah kalimat d adalah kalimat berikutnya dari kalimat q .

Selama proses pelatihan, setengah dari *input* tersebut adalah pasangan kalimat di mana kalimat kedua adalah kalimat berikutnya, dan setengah lainnya adalah kalimat yang diambil secara acak dari korpus sebagai kalimat kedua. Gambar 3.12 mengilustrasikan contoh dari tugas *next sentence prediction*.



Gambar 3.12: Ilustrasi *next sentence prediction* pada BERT. Model diminta untuk memprediksi apakah kalimat kedua adalah kalimat berikutnya dari kalimat pertama.

3.3.3 BERT untuk Bahasa Indonesia (IndoBERT)

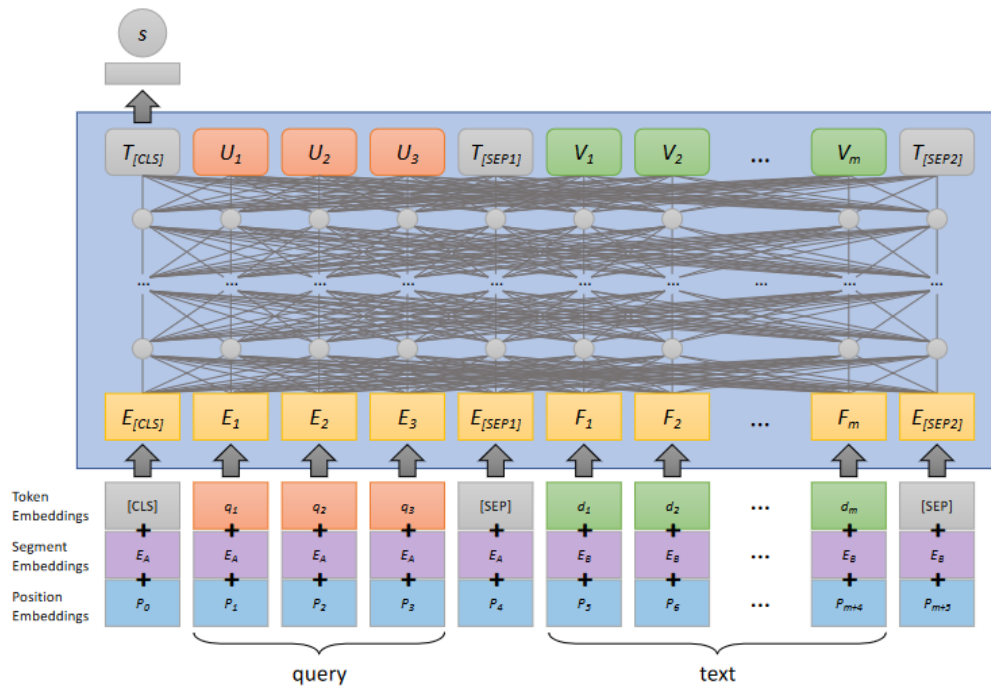
Pada penelitian ini, BERT yang digunakan adalah BERT untuk bahasa Indonesia yang sudah dilakukan *pre-training* sebelumnya oleh Koto, Rahimi, Lau, dan Baldwin (2020).

Pre-training BERT untuk bahasa Indonesia dilakukan dengan menggunakan korpus Wikipedia bahasa Indonesia dengan 74 Juta kata, artikel berita dari Kompas, Tempo, dan Liputan6 dengan 55 Juta kata, dan korpus web bahasa Indonesia dengan 90 Juta kata. model IndoBERT dilatih selama 2.4 Juta iterasi (180 epoch).

3.3.4 Penggunaan BERT untuk Pemeringkatan Teks

Bagian berikut akan menjelaskan penggunaan BERT untuk pemeringkatan teks. terdapat dua arsitektur yang digunakan, yaitu BERT_{CAT} dan BERT_{DOT} yang masing-masing akan dijelaskan pada Subbab 3.3.4.1 dan Subbab 3.3.4.2.

3.3.4.1 BERT_{CAT}



Gambar 3.13: BERT_{CAT} mengambil kueri dan kandidat teks yang akan diberi skor sebagai *input* dan menggunakan BERT untuk klasifikasi relevansi. Penjumlahan elemen-wise dari token, *segment*, dan *positional embeddings* membentuk representasi vektor *input*. Setiap token input memiliki vektor kontekstual sebagai *output* model BERT. *Linear layer* menerima representasi akhir token [CLS] dan menghasilkan skor relevansi teks terkait dengan kueri (Lin et al., 2020).

Salah satu cara penggunaan BERT untuk pemeringkatan Teks adalah dengan menggunakan BERT pada model untuk melakukan *soft classification* nilai relevansi dari pasangan (kueri, teks). Dengan kata lain, skor relevansi dari pasangan (kueri, teks) adalah

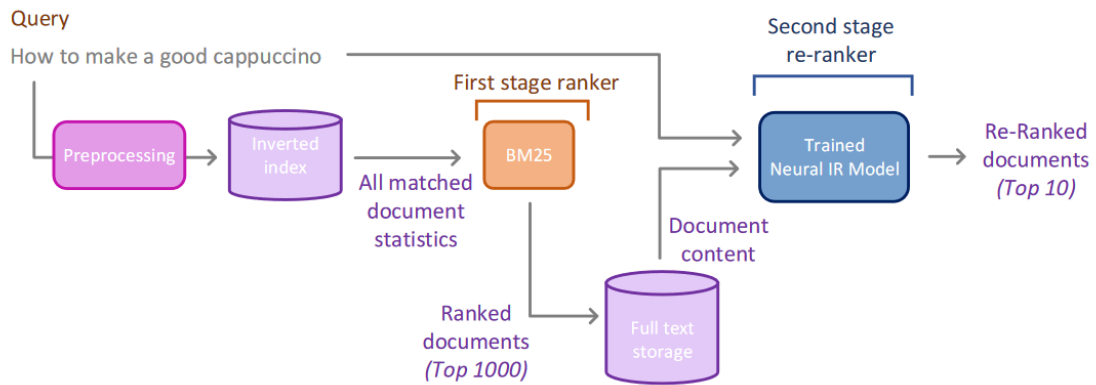
probabilitas bahwa teks tersebut relevan dengan kueri seperti yang ditunjukkan pada Persamaan 3.28 hingga Persamaan 3.29.

$$\text{score}(q, d) = P(\text{relevance} = 1 | q, d) = \sigma \left(\mathbf{h}_{[\text{CLS}]} \mathbf{W}^{\text{CLS}} + \mathbf{b}^{\text{CLS}} \right), \quad (3.28)$$

$$\mathbf{h}_{[\text{CLS}]} = \text{BERT}([[\text{CLS}], q, [\text{SEP}], d, [\text{SEP}]])_{[\text{CLS}]} \in \mathbb{R}^{d_{\text{token}}}, \quad (3.29)$$

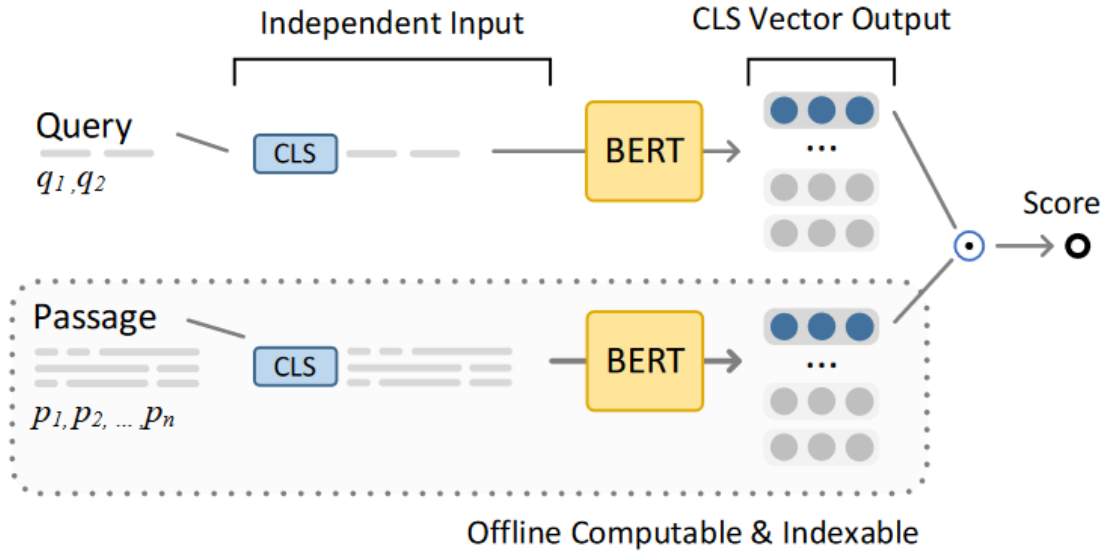
dengan $\mathbf{W}^{\text{CLS}} \in \mathbb{R}^{d_{\text{token}} \times 1}$ dan $\mathbf{b}^{\text{CLS}} \in \mathbb{R}$ adalah matriks bobot dan bias yang digunakan untuk melakukan *classification*, dan σ adalah fungsi sigmoid.

Untuk suatu kueri q dan kumpulan teks $D = \{d_1, d_2, \dots, d_n\}$, perlu dilakukan perhitungan skor relevansi untuk setiap pasangan (q, d_i) dengan $i = 1, \dots, n$ sebelum dilakukan pemeringkatan. hal ini menjadi masalah untuk jumlah dokumen yang besar karena setiap perhitungan skor relevansi dengan BERT_{CAT} membutuhkan waktu yang lama. Oleh karena itu, BERT_{CAT} biasanya digunakan sebagai *reranker* dari sistem pemeringkatan teks. teks yang akan diberikan skor relevansinya oleh BERT_{CAT} adalah teks yang sudah dipilih oleh sistem pemeringkatan teks yang lebih efisien seperti BM25 – biasanya $k = 100, 1000$ teks teratas dipilih oleh BM25. Gambar 3.14 mengilustrasikan arsitektur *retrieve* dan *rerank*.



Gambar 3.14: Arsitektur *retrieve* and *rerank*. *First-stage retrieval* dilakukan oleh BM25 dan *reranking* dilakukan oleh model *scoring* yang lebih kompleks seperti BERT_{CAT} (Hofstätter et al., 2021).

3.3.4.2 BERT_{DOT}



Gambar 3.15: BERT_{DOT} memetakan kueri dan kandidat teks ke dalam ruang vektor yang sama dan menghitung skor relevansi dengan melakukan *dot product* antara vektor representasi kontekstual dari kueri dan teks (Hofstätter et al., 2021).

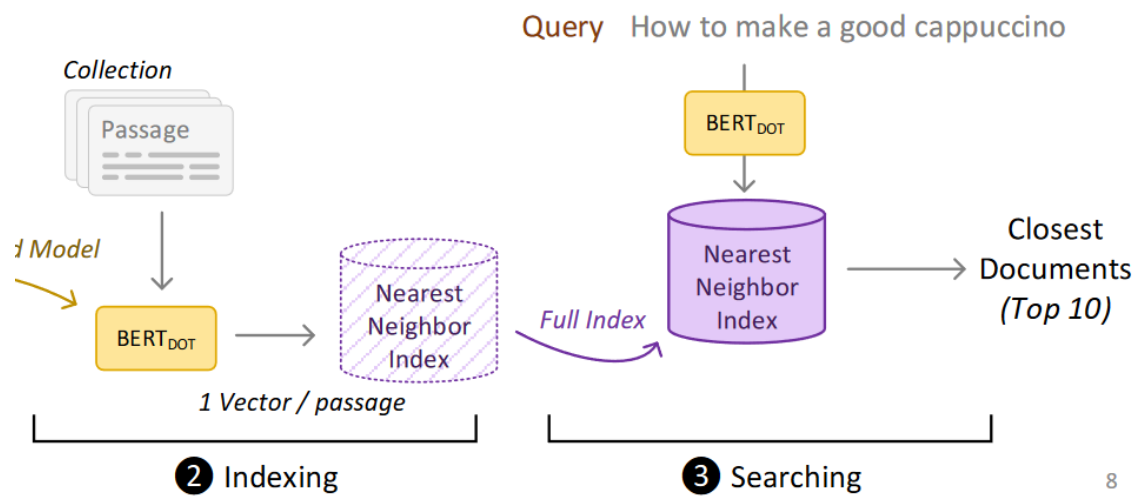
Berbeda dengan BERT_{CAT}, BERT_{DOT} tidak melakukan *soft classification* untuk setiap pasangan (kueri, teks). BERT_{DOT} menghitung skor relevansi dari pasangan (kueri, teks) dengan melakukan *dot product* antara vektor representasi kontekstual dari kueri dan teks seperti yang ditunjukkan pada Persamaan 3.30 hingga Persamaan 3.32.

$$\mathbf{q}_{[\text{CLS}]} = \text{BERT}([[\text{CLS}], q, [\text{SEP}]])_{[\text{CLS}]} \in \mathbb{R}^{d_{\text{token}}}, \quad (3.30)$$

$$\mathbf{d}_{[\text{CLS}]} = \text{BERT}([[\text{CLS}], d, [\text{SEP}]])_{[\text{CLS}]} \in \mathbb{R}^{d_{\text{token}}}, \quad (3.31)$$

$$\text{score}(q, d) = \mathbf{q}_{[\text{CLS}]} \mathbf{d}_{[\text{CLS}]}^T \in \mathbb{R}. \quad (3.32)$$

Salah satu kelebihan dari BERT_{DOT} adalah vektor representasi dari setiap teks dapat dihitung terlebih dahulu dan disimpan dalam memori. Akibatnya, kita hanya perlu menghitung vektor dari kueri dan nilai *dot product* untuk setiap teksnya ketika akan dilakukan pemeringkatan – yang tentunya lebih efisien secara komputasi. Hal ini membuat BERT_{DOT} lebih cepat daripada BERT_{CAT} untuk melakukan pemeringkatan. Gambar 3.16 mengilustrasikan arsitektur pemeringkatan dengan BERT_{DOT}.



8

Gambar 3.16: Arsitektur pemeringkatan dengan BERT_{DOT}. Vektor representasi dari setiap teks dapat diindeks terlebih dahulu dan disimpan dalam memori (Hofstätter et al., 2021).

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 pada masalah pemeringkatan teks. Subbab 4.1 menjelaskan mengenai spesifikasi perangkat keras dan perangkat lunak yang digunakan dalam penelitian. Selanjutnya, Subbab 4.2 menjelaskan mengenai tahapan simulasi yang dilakukan dalam penelitian. Informasi mengenai *dataset* latih dan uji dijelaskan pada Subbab 4.3. Subbab 4.4 menjelaskan lebih detail mengenai arsitektur model BERT, fungsi loss, serta konfigurasi *hyperparameter* yang digunakan dalam proses *fine tuning* model BERT. Terakhir, Subbab 4.5 menjelaskan mengenai evaluasi hasil *fine tuning* model BERT untuk pemeringkatan teks.

4.1 Spesifikasi Mesin dan Perangkat Lunak

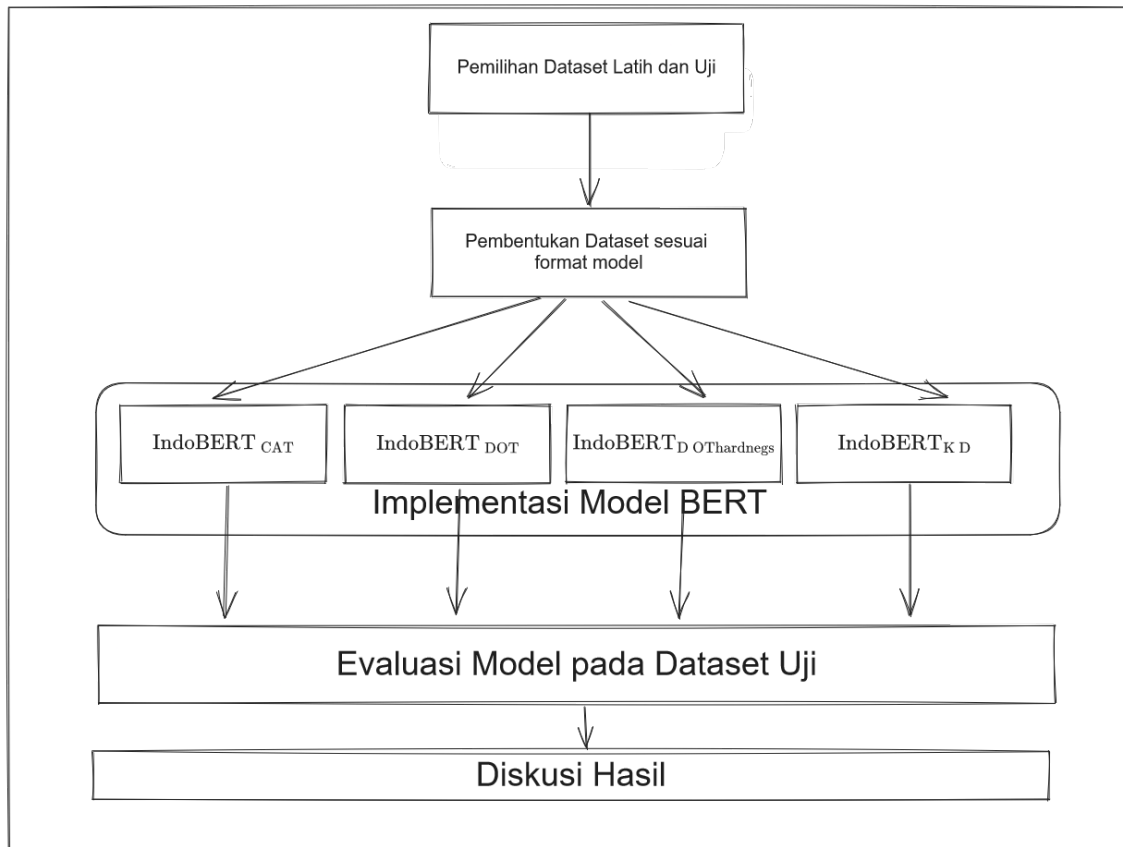
Proses *fine tuning* model BERT untuk pemeringkatan teks dilakukan menggunakan mesin dan perangkat lunak yang tertera pada Tabel 4.1.

Tabel 4.1: Spesifikasi perangkat lunak yang digunakan pada penelitian ini.

CPU	AMD Ryzen 9 5950X 32-Core Processor
GPU	NVIDIA GeForce RTX 4090 24GB
Memori	64GB
Sistem Operasi	Ubuntu 20.04.2 LTS
Perangkat Lunak Pemrograman	Visual Studio Code 1.84.2
Bahasa Pemrograman	Python 3.8
Pustaka yang Digunakan	sentence-transformers 2.2.2 transformers 4.35.1 beir 2.0.0 gdown 4.7.1 torch 2.0.1+cu117

4.2 Tahapan Simulasi

Gambar 4.1 menunjukkan tahapan simulasi yang dilakukan dalam penelitian ini.



Gambar 4.1: Diagram Simulasi

Simulasi diawali dengan pengambilan data. Data yang digunakan adalah data pada penelitian Bonifacio, Campiotti, de Alencar Lotufo, dan Nogueira (2021), sebagai *dataset* latih, dan data pada penelitian X. Zhang, Ma, Shi, dan Lin (2021), X. Zhang et al. (2023) sebagai *dataset* uji. *Dataset* latih tidak dapat digunakan langsung untuk melatih model-model tersebut. Untuk setiap modelnya, diperlukan transformasi untuk mengubah bentuk dari *dataset* latih sehingga sesuai dengan formatnya. Transformasi *dataset* latih dan *hyperparameter* dari model akan dibahas lebih lanjut pada bagian Subbab 4.4. Selanjutnya, proses implementasi dan pelatihan model dilakukan. Setelah itu, akan dilakukan evaluasi setiap model pada *dataset* uji. Terdapat satu model BM25 sebagai *baseline* untuk membandingkan hasil evaluasi dari model-model yang dilatih. Terakhir, terdapat diskusi mengenai hasil evaluasi tersebut.

4.3 Data

Penelitian ini menggunakan satu *dataset* latih Mmarco *train set* bahasa Indonesia (Bonifacio et al., 2021) dan tiga *dataset* uji, yaitu Mmarco *dev set* bahasa Indonesia, MrTyDi Test set Indonesia (X. Zhang et al., 2021), dan Miracl *dev set* bahasa Indonesia (X. Zhang et al., 2023). *Dataset* Miracl dan MrTyDi dipilih sebagai uji kemampuan *out-of-distribution* dari model yang dihasilkan. Setiap *dataset* terdiri dari 3 *file*, yaitu *file* kueri, *file* korpus dan *file* *judgements* yang telah dijelaskan pada Subbab 2.1.1. Tabel 4.2 menunjukkan informasi mengenai jumlah entri dari *file* kueri, *file* korpus, dan *file* *judgements* dari setiap *dataset* yang digunakan dalam penelitian ini.

Tabel 4.2: Tabel Informasi untuk Setiap *Dataset*. Kolom *Korpus* menunjukkan jumlah entri pada *file* korpus, kolom *Kueri* menunjukkan jumlah entri pada *file* kueri, dan kolom *Judgements* menunjukkan jumlah entri pada *file* *judgements* (pasangan kueri dan teks dengan nilai relevansi).

Sumber: (Tsai et al., 2019)

Dataset	Korpus	Kueri	Judgements
Mmarco train set	8,841,823	1,010,916	532,761
Mmarco dev set	8,841,823	1,010,916	7,437
Mrtydi test set	1,469,399	829	961
Miracl dev set	1,446,315	960	9,668

Setiap *judgements* pada *dataset* adalah *judgements* biner, yaitu bernilai 1 jika teks tersebut relevan dengan kueri dan 0 jika tidak relevan dengan kueri. Tabel 4.3 hingga Tabel 4.5 menunjukkan contoh dari *file* korpus, *file* kueri, dan *file* *judgements* pada *dataset* Miracl *dev set*.

Tabel 4.3: Potongan *file* korpus *dataset* Miracl.

id	title	text
1342516#1	Colobothea biguttata	Larva kumbang ini biasanya mengebor ke dalam kayu dan dapat menyebabkan kerusakan pada batang kayu hidup atau kayu yang telah ditebang.
1342517#0	Ichthyodes rufipes	Ichthyodes rufipes adalah spesies kumbang tanduk panjang yang berasal dari famili Cerambycidae. Spesies ini juga merupakan bagian dari genus Ichthyodes, ordo Coleoptera, kelas Insecta, filum Arthropoda, dan kingdom Animalia.

Tabel 4.4: Potongan *file* kueri *dataset* Miracl.

id	text
3	Dimana James Hepburn meninggal?
4	Dimana Jamie Richard Vardy lahir?
11	berapakah luas pulau Flores?
17	Siapakah yang menulis Candy Candy?
19	Apakah karya tulis Irma Hardisurya yang pertama?

Tabel 4.5: Potongan *file* judgements *dataset* Miracl.

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

4.4 *Fine Tuning* Model BERT

Bagian ini menjelaskan mengenai konfigurasi *hyperparameter* yang digunakan pada setiap model yang dikerjakan pada penelitian ini. Terdapat empat model yang dikerjakan pada penelitian ini, yaitu $\text{IndoBERT}_{\text{CAT}}$, $\text{IndoBERT}_{\text{DOT}}$, $\text{IndoBERT}_{\text{DOTHardnegs}}$, dan $\text{IndoBERT}_{\text{DOTKD}}$. Keempat model tersebut merupakan model BERT yang dilatih dengan menggunakan *dataset* Mmarco *train set* dengan prosedur yang berbeda-beda.

4.4.1 $\text{IndoBERT}_{\text{CAT}}$

Pada model $\text{IndoBERT}_{\text{CAT}}$, arsitektur BERT_{CAT} (lihat Subbab 3.3.4.1) digunakan untuk melakukan pemeringkatan teks. Proses pelatihan model menggunakan *dataset* yang digunakan berasal dari Mmarco *train set* dengan format (q, d, r) dengan q adalah kueri, d adalah teks, dan r adalah relevansi teks d terhadap kueri q . Pelatihan yang dilakukan seperti melakukan klasifikasi relevansi teks terhadap kueri. Perlu dicatat bahwa tidak ada contoh $r = 0$ dalam *dataset* Mmarco *train set* (lihat Subbab 2.1.1).

Untuk membentuk data latih dengan penilaian $r = 0$, pasangan $(q, d', 0)$ ditambahkan dengan d' sebagai teks acak yang tidak relevan dengan kueri q . *Dataset* yang telah dibuat

terdiri dari 500 ribu pasangan (q, d, r) dengan rasio 1:1 antara $r = 1$ dan $r = 0$. Potongan *dataset* yang digunakan untuk pelatihan model IndoBERT_{CAT} dapat ditemukan pada Tabel 4.6.

Tabel 4.6: Potongan *dataset* yang digunakan untuk pelatihan model IndoBERT_{CAT}.

Kueri	Teks	Relevansi
Berapa banyak kalori sehari yang hilang saat menyusui?	Tidak hanya menyusui lebih baik untuk bayi, namun penelitian juga mengatakan itu lebih baik bagi ibu. Menyusui membakar rata-rata 500 kalori sehari, dengan kisaran khas antara 200 hingga 600 kalori yang terbakar sehari. Diperkirakan produksi 1 oz. ASI membakar 20 kalori. Jumlah kalori yang terbakar tergantung pada seberapa banyak bayi makan. Menyusui kembar membakar dua kali lebih banyak daripada memberi makan hanya satu bayi. Dengan anak kembar, ibu mereka membakar 1000 kalori per hari. Membakar 500 kalori ekstra sehari akan menghasilkan satu pon penurunan berat badan mingguan.	1
Karakteristik iklim utama hutan hujan tropis	Kacang kola adalah buah dari pohon kola, genus (Cola) pohon yang berasal dari hutan hujan tropis Afrika.	0

Konfigurasi *hyperparameter* selama pelatihan model indoBERT_{CAT} dapat dilihat pada Tabel 4.7.

Tabel 4.7: *Hyperparameter* yang digunakan untuk *fine tuning* IndoBERT_{CAT}.

Parameter	Nilai
Model pralatih	indolem/indobert-base-uncased
Total data	500,000
<i>Batch size</i>	32
Total iterasi	78125 (5 epochs)
<i>Optimizer</i>	Adam dengan $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$
<i>Learning rate</i>	2e-5
<i>Learning rate warmup</i>	Linear selama 10% dari total iterasi
Fungsi loss	<i>Binary cross entropy</i>

4.4.2 IndoBERT_{DOT}

Pada model IndoBERT_{DOT}, arsitektur BERT_{DOT} (lihat Subbab 3.3.4.2) digunakan untuk melakukan pemeringkatan teks. fungsi loss yang digunakan untuk pelatihan model IndoBERT_{DOT} adalah *N-pair loss*. Untuk kueri q , teks relevan d^+ , dan kumpulan teks tidak relevan $\{d_i^-\}_{i=1}^{N-1}$ terhadap kueri q , *N-pair loss* dihitung sebagai berikut:

$$L(q, d^+, \{d_i^-\}_{i=1}^{N-1}) = -\log \frac{\exp(\mathbf{h}_q^\top \mathbf{h}_d^+)}{\exp(\mathbf{h}_q^\top \mathbf{h}_d^+) + \sum_{i=1}^{N-1} \exp(\mathbf{h}_q^\top \mathbf{h}_i^-)}, \quad (4.1)$$

dengan keterangan sebagai berikut:

$$\mathbf{h}_q = \text{IndoBERT}_{\text{DOT}}([[\text{CLS}], q, [\text{SEP}]])_{[\text{CLS}]}$$

$$\mathbf{h}_d^+ = \text{IndoBERT}_{\text{DOT}}([[\text{CLS}], d^+, [\text{SEP}]])_{[\text{CLS}]}$$

$$\mathbf{h}_i^- = \text{IndoBERT}_{\text{DOT}}([[\text{CLS}], d_i^-, [\text{SEP}]])_{[\text{CLS}]}$$

Pasangan (q, d^+) diambil langsung dari *file judgements* pada *Mmarco train set*. Kumpulan teks tak relevan $\{d_i^-\}_{i=1}^{N-1}$ dibentuk dengan menggunakan teks d pada *data point* yang berbeda pada *batch* yang sama. Nilai N pada *N-pair loss* adalah ukuran *batch* yang digunakan selama pelatihan model. Metode pemilihan teks negatif ini disebut dengan *in-batch negative sampling* (Karpukhin et al., 2020). Pada penelitian ini, digunakan seluruh *datapoint* pada *file judgements* *Mmarco train set* – 532,761 *data point* – untuk

membentuk *dataset* latih model IndoBERT_{DOT}.

Konfigurasi *hyperparameter* selama pelatihan model indoBERT_{DOT} dapat dilihat pada Tabel 4.8.

Tabel 4.8: *Hyperparameter* yang digunakan untuk *fine tuning* IndoBERT_{DOT}.

Parameter	Nilai
Model pralatih	indolem/indobert-base-uncased
Total data	532,761
<i>Batch size</i>	32
Total iterasi	83243 (5 epochs)
<i>Optimizer</i>	Adam dengan $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$
<i>Learning rate</i>	$2e-5$
<i>Learning rate warmup</i>	Linear selama 10% dari total iterasi
Fungsi <i>loss</i>	<i>N-pair loss</i>

4.4.3 IndoBERT_{DOThardnegs}

Pada IndoBERT_{DOThardnegs}, arsitektur BERT_{DOT} digunakan untuk melakukan pemerinkatan teks. Fungsi *loss* yang digunakan untuk pelatihan model IndoBERT_{DOThardnegs} adalah *N-pair loss* seperti IndoBERT_{DOT}. Perbedaan utama antara IndoBERT_{DOT} dan IndoBERT_{DOThardnegs} pada metode pemilihan teks negatif. Pada IndoBERT_{DOThardnegs}, teks negatif sudah terlebih dahulu dipilih dengan kriteria bahwa teks tersebut adalah teks yang tidak relevan dengan kueri q , tetapi pemerinkatan dengan BM25 berada di 100 teratas. Dengan kata lain, teks negatif adalah teks yang sulit dibedakan dengan teks positif ketika menggunakan BM25. Teks *hard negative* ini diberikan oleh suatu *file* dan Tabel 4.9 menunjukkan contoh dari teks *file* tersebut. Nilai N yang dipilih pada penelitian ini adalah $N = 5$, dan jumlah data yang digunakan adalah 502.939 *data point*.

Tabel 4.9: Potongan *file* sentence-transformers/msmarco-hard-negatives. kolom *qid* berisikan id dari kueri, kolom *positive* adalah id teks positif, dan kolom *hard negative* adalah id teks yang sulit dibedakan dengan teks positif menggunakan BM25.

qid	Positive	Hard Negative
634	1668221	{ "bm25": [830424, 1985345, 1153265, 1638798, 2866545] }
492	875	{ "bm25": [1147445, 4859158, 7980570, 3409840, 6889336] }

Konfigurasi *hyperparameter* selama pelatihan model indoBERT_{DOThardnegs} dapat dilihat pada Tabel 4.10.

Tabel 4.10: *Hyperparameter yang digunakan untuk fine tuning IndoBERT_{DOTHardnegs}.*

Parameter	Nilai
Model prlatih	indolem/indobert-base-uncased
Total data	502,939
Batch Size	32
Total Iterasi	78585 (5 epochs)
Optimizer	Adam dengan $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$
Learning rate	2e-5
Learning rate warmup	Linear selama 10% dari total iterasi
Fungsi loss	<i>N-pair loss</i>

4.4.4 IndoBERT_{DOTKD}

IndoBERT_{DOTKD} dilatih dengan menggunakan prinsip *knowledge distillation*, yaitu proses *transfer* pengetahuan dari model yang sudah dilatih dengan baik (guru) ke model yang belum dilatih (murid). Model yang digunakan sebagai guru adalah model bahasa Inggris yang sudah dilatih dengan baik untuk melakukan pemeringkatan teks. Model Murid yang dapat dipilih adalah model prlatih BERT multibahasa – model yang *pre-training*-nya dilakukan pada korpus multibahasa seperti mBERT (*multilingual BERT*). Pemetaan vektor dari model murid akan di-*align* dengan Pemetaan vektor dari model guru dengan fungsi *loss* berikut:

$$L(s_i, t_i) = ((\| M(s_i) - \hat{M}(s_i) \|)^2 + (\| M(s_i) - \hat{M}(t_i) \|)^2), \quad (4.2)$$

dengan keterangan sebagai berikut:

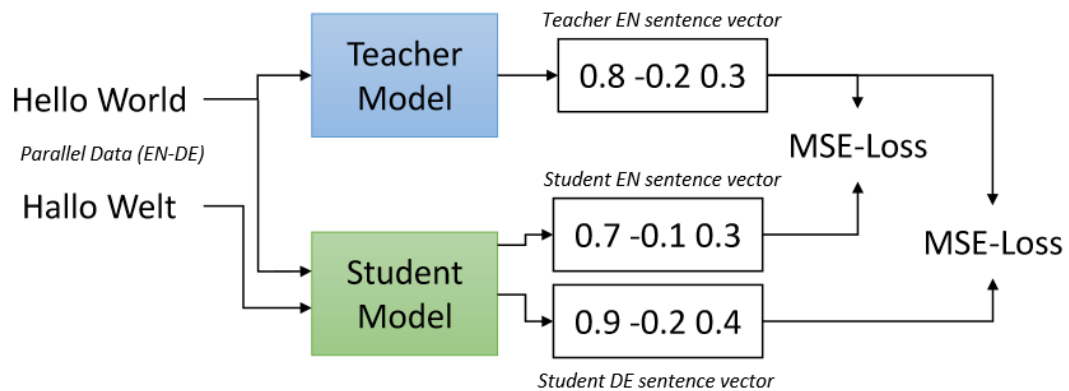
M = pemetaan vektor oleh model guru,

\hat{M} = pemetaan vektor oleh model murid,

s_i = teks sumber (bahasa Inggris),

t_i = teks target (bahasa Indonesia).

Gambar 4.2 menunjukkan ilustrasi dari proses pelatihan dengan *knowledge distillation*. 500 ribu kueri dan 500 ribu korpus dari Mmarco *train set* Indonesia di-*align* dengan terjemahannya seperti yang ditunjukkan pada Tabel 4.11.



Gambar 4.2: Ilustrasi dari pelatihan model IndoBERT_{DOTKD} dengan *knowledge distillation*. Kalimat paralel diberikan sebagai *input* pada model guru dan model murid. vektor yang dihasilkan oleh model guru dan model murid di-align menggunakan fungsi *loss mean squared error*.

Sumber: (Reimers & Gurevych, 2020)

Tabel 4.11: Potongan dari *dataset* yang digunakan untuk pelatihan model IndoBERT_{KD}.

Sumber: carles-undergrad-thesis/en-id-parallel-sentences

text_en	txt_id
<i>Defining alcoholism as a disease is associated with Jellinek</i>	Mendefinisikan alkoholisme sebagai penyakit dikaitkan dengan Jellinek
<i>ECT is a treatment that is used for</i>	ECT adalah pengobatan yang digunakan untuk
<i>Ebolavirus is an enveloped virus, which means</i>	Ebolavirus adalah virus yang diselimuti, yang berarti
<i>How much does Cambridge Manor cost per month</i>	Berapa biaya Cambridge Manor per bulan?

Konfigurasi *hyperparameter* selama pelatihan model indoBERT_{KD} dapat dilihat pada Tabel 4.12.

Tabel 4.12: *Hyperparameter* yang digunakan untuk *fine tuning* IndoBERT_{DOTKD}.

Parameter	Nilai
Model guru	sentence-transformers/msmarco-bert-base-dot-v5
Model murid	bert-base-multilingual-uncased
Total data	1,000,000
<i>Batch Size</i>	64
Total Iterasi	78125 (5 epochs)
<i>Optimizer</i>	Adam dengan $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$
<i>Learning rate</i>	2e-5
<i>Learning rate warmup</i>	Linear selama 10% dari total iterasi
Fungsi <i>loss</i>	<i>Mean squared error</i>

4.5 Evaluasi Model

Subbab ini membahas perihal hasil dari *Fine tuning* dan evaluasi dari keempat model. BM25 digunakan sebagai *baseline* untuk membandingkan evaluasi setiap model. BM25 yang digunakan adalah *software Elastic search* (<https://www.elastic.co/elasticsearch/>) untuk bahasa Indonesia dengan konfigurasi parameter *default*.

Metrik yang digunakan pada *datasets* Mmarco *dev set* dan MrTyDi *test set* adalah *Reciprocal Rank* pada 10 teks teratas (RR@10) dan *Recall* pada 1000 teks teratas (R@1000). Berbeda dengan mMarco dan MrTyDi, Metrik yang digunakan pada *dataset* Miracl *dev set* adalah *Normalized Discounted Cumulative Gain* pada 10 teks teratas (NDCG@10) dan *Recall* pada 1000 teks teratas (R@1000). Perhitungan metrik lainnya untuk setiap model dapat dilihat pada bagian lampiran.

4.5.1 Evaluasi IndoBERT_{CAT}

Tabel 4.13: Evaluasi model IndoBERT_{CAT} pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*.

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	RR@10	R@1000	RR@10	R@1000	NDCG@10	R@1000
BM25	.114	.642	.279	.858	.391	.971
BM25+IndoBERT _{CAT}	.181	.642	.447	.858	.455	.971

Tabel 4.13 menunjukkan evaluasi dan perbandingan antara model IndoBERT_{CAT} dan BM25. Perlu diingat kembali bahwa model IndoBERT_{CAT} digunakan sebagai *reranker* (lihat Subbab 3.3.4.1) untuk memperbaiki hasil dari BM25. Efeknya nilai *Recall* dari model IndoBERT_{CAT} sama dengan nilai *Recall* dari BM25. Namun, nilai *Reciprocal Rank* dari model IndoBERT_{CAT} lebih tinggi dari BM25. Hal ini menunjukkan bahwa model IndoBERT_{CAT} lebih baik dalam memberikan urutan teks yang direkomendasikan oleh BM25.

nilai RR@10 dari model IndoBERT_{CAT} pada dataset Mmarco *dev set* meningkat sebesar .67 poin (58%) dibandingkan dengan BM25. Peningkatan yang sama juga terjadi pada nilai RR@10 MrTyDi *test set* dan nilai NDCG@10 Miracl *dev set* dengan peningkatan sebesar .168 poin (60%) dan .064 poin (16%) masing-masingnya.

4.5.2 Evaluasi IndoBERT_{DOT}

Tabel 4.14: Evaluasi model IndoBERT_{DOT} pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*.

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	RR@10	R@1000	RR@10	R@1000	NCDG@10	R@1000
BM25	.114	.642	.279	.858	.391	.971
IndoBERT _{DOT}	.192	.847	.378	.936	.355	.920

Tabel 4.14 menunjukkan evaluasi dan perbandingan antara model IndoBERT_{DOT} dan BM25. Peningkatan terjadi pada nilai RR@10 dan R@1000 pada *dataset* Mmarco *dev set* dan MrTyDi *test set*. Nilai RR@10 pada Mmarco *dev set* meningkat sebesar .078 poin (68%) dan pada MrTyDi *test set* sebesar .099 poin (35%). Nilai R@1000 pada Mmarco *dev set* meningkat sebesar .205 poin (32%) dan pada MrTyDi *test set* sebesar .078 poin (9%). Sementara itu, nilai NDCG@10 pada Miracl *dev set* menurun sebesar .036 poin (-9%) dan nilai R@1000 juga menurun sebesar .051 poin (-5%).

4.5.3 Evaluasi IndoBERT_{DOTHardnegs}

Tabel 4.15: Evaluasi model IndoBERT_{DOTHardnegs} pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*.

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	RR@10	R@1000	RR@10	R@1000	NCDG@10	R@1000
BM25	.114	.642	.279	.858	.391	.971
IndoBERT _{DOTHardnegs}	.232	.847	.471	.921	.397	.898

Tabel 4.15 menunjukkan evaluasi dan perbandingan antara model IndoBERT_{DOTHardnegs} dan BM25. Peningkatan terjadi pada nilai RR@10 dan R@1000 pada *dataset* Mmarco *dev set* dan MrTyDi *test set*. Nilai RR@10 pada Mmarco *dev set* meningkat sebesar .118 poin (103%) dan pada MrTyDi *test set* sebesar .192 poin (69%). Nilai R@1000 pada Mmarco *dev set* juga meningkat sebesar .205 poin (32%) dan pada MrTyDi *test set* sebesar .063 poin (7%). Sementara itu, nilai NDCG@10 pada Miracl *dev set* menurun sebesar .006 poin (-1%) dan nilai R@1000 juga menurun sebesar .073 poin (-8%).

4.5.4 Evaluasi IndoBERT_{KD}

Tabel 4.16: Evaluasi model IndoBERT_{KD} pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*.

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	RR@10	R@1000	RR@10	R@1000	NCDG@10	R@1000
BM25	.114	.642	.279	.858	.391	.971
IndoBERT _{DOTKD}	.235	.867	.393	.882	.374	.871

Tabel 4.16 menunjukkan evaluasi dan perbandingan antara model IndoBERT_{DOTKD} dengan BM25. Peningkatan Terjadi pada nilai RR@10 dan R@1000 pada *dataset* Mmarco *dev set* dan MrTyDi *test set*. Nilai RR@10 pada Mmarco *dev set* meningkat sebesar .121 poin (106%) dan pada MrTyDi *test set* sebesar .114 poin (41%). Nilai R@1000 pada Mmarco *dev set* juga meningkat sebesar .225 poin (37%) dan pada MrTyDi *test set* sebesar .024 poin (3%). Sementara itu, nilai NDCG@10 pada Miracl *dev set* menurun sebesar .017 poin (-4%) dan nilai R@1000 juga menurun sebesar .100 poin (-10%).

4.6 Diskusi Hasil

Tabel 4.17: Evaluasi dari model $\text{IndoBERT}_{\text{CAT}}$, $\text{IndoBERT}_{\text{DOT}}$, $\text{IndoBERT}_{\text{DOTHardnegs}}$, dan $\text{IndoBERT}_{\text{DOTKD}}$ pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*.

Model	Mmarco Dev		MrTyDi Test		Miracl Dev	
	RR@10	R@1000	RR@10	R@1000	NCDG@10	R@1000
BM25	.114	.642	.279	.858	.391	.971
BM25+ $\text{IndoBERT}_{\text{CAT}}$.181	.642	.447	.858	.455	.971
$\text{IndoBERT}_{\text{DOT}}$.192	.847	.378	.936	.355	.920
$\text{IndoBERT}_{\text{DOTHardnegs}}$.232	.847	.471	.921	.397	.898
$\text{IndoBERT}_{\text{DOTKD}}$.235	.867	.393	.882	.374	.871

Tabel 4.17 menunjukkan evaluasi dari model $\text{IndoBERT}_{\text{CAT}}$, $\text{IndoBERT}_{\text{DOT}}$, $\text{IndoBERT}_{\text{DOTHardnegs}}$, dan $\text{IndoBERT}_{\text{DOTKD}}$ pada *dataset* Mmarco *dev set*, MrTyDi *test set*, dan Miracl *dev set*. Perhatikan bahwa model $\text{IndoBERT}_{\text{CAT}}$ adalah model yang *robust* dibandingkan dengan model lainnya – Peningkatan metrik terjadi disetiap *dataset* yang digunakan. Hal ini dapat dikaitkan dengan fakta bahwa model dengan arsitektur BERT_{CAT} dapat mengaitkan kata-kata pada kueri dengan kata-kata pada teks ketika proses pemberian skor. Namun, meskipun $\text{IndoBERT}_{\text{CAT}}$ adalah model yang paling *robust*, peningkatan setiap metrik tidaklah terlalu besar seperti model yang berasitekturkan BERT_{DOT} . Dugaan yang dapat penulis berikan adalah proses pelatihan tersebut membutuhkan waktu yang lebih lama untuk menghasilkan fungsi skoring yang baik karena cara melatih model BERT_{CAT} yang berupa klasifikasi relevansi antara kueri dan teks. Pelatihan yang dilakukan pada model $\text{IndoBERT}_{\text{CAT}}$ memiliki kelebihan dan kekurangan tersendiri. Kelebihan yang didapat adalah nilai skor antar kueri dan teks memiliki makna, meskipun berdiri sendiri. Nilai skor antara kueri dan dokumen menunjukkan nilai relevansinya antara kueri dan teks tersebut, hal ini kontras dengan model BERT_{DOT} yang hanya menghasilkan skor antara kueri dan teks, tanpa memiliki makna. Di lain sisi, kekurangannya adalah tugas klasifikasi relevansi tidak secara langsung melatih dengan tujuan yang diharapkan. Pada proses pemeringkatan teks, skor antara kueri dan teks tidak perlu memiliki makna, yang diinginkan hanyalah skor antara kueri dan teks yang relevan lebih tinggi dibandingkan skor antara kueri dan teks tidak relevan.

Berbeda dengan arsitektur BERT_{CAT} , arsitektur BERT_{DOT} dilatih dengan tujuan yang

diharapkan secara langsung dengan N -pair loss. Hal ini dapat dilihat dari hasil evaluasi pada Tabel 4.17 dimana model IndoBERT_{DOT} dan IndoBERT_{DOTHardnegs} memiliki peningkatan yang lebih besar dibandingkan model IndoBERT_{CAT} pada Mmarco dev set – in-domain dataset. Pemilihan *hard negative* teks pada model IndoBERT_{DOTHardnegs} juga meningkatkan performa model. Hal ini dapat dilihat pada evaluasi di dataset lainnya.

Skor yang dihasilkan oleh model IndoBERT_{DOT} dan IndoBERT_{DOTHardnegs} tidak memiliki makna jika berdiri sendiri. Skor antara kueri dan teks hanya digunakan untuk membandingkan skor antara kueri dan teks lainnya. Skor antara kueri dan teks tidak dapat digunakan untuk mengetahui nilai relevansi antara kueri dan teks tersebut.

Model IndoBERT_{DOTKD} memiliki performa yang terbaik pada in-domain dataset Mmarco dev set. Namun performa tersebut tidaklah sebaik model IndoBERT_{DOTHardnegs} pada out-of-domain dataset MrTyDi test set dan Miracl dev set. Pelatihan dengan meng-align pemetaan vektor antara model murid dan model guru tidak baik untuk mengeneralisasi teks dan kueri yang tidak ada pada dataset pelatihan.

Tabel 4.18: Benchmark model BERT_{DOT} dan BERT_{CAT}, dan BM25 pada dataset Mmarco dev set. Latensi dan memori diukur pada hardware yang sama dengan yang digunakan pada pelatihan model.

Model	Latensi (ms)	Memori(MB)
BM25 (Elastic Search)	6.55 (CPU)	800
BERT _{DOT}	9.9 (GPU)	3072
BM25+BERT _{CAT}	242 (GPU)	800

Tabel 4.18 menunjukkan latensi dan memori yang digunakan oleh model berasitekturkan BERT_{DOT}, BERT_{CAT}, dan BM25. Latensi dan memori diukur pada hardware yang sama dengan yang digunakan pada pelatihan model. Model berasitekturkan BERT_{DOT} Memiliki latensi yang hampir sama dengan BM25 dan Latensi yang lebih baik dibandingkan dengan model BERT_{CAT}. Hal ini dapat disebabkan model fungsi skor pada BERT_{DOT} hanyalah berupa perkalian titik dan vektor teks dapat dihitung (*indexing*) terlebih dahulu sebelum melakukan pemeringkatan teks. Bandingkan dengan model BERT_{CAT} yang memiliki fungsi skor yang kompleks dan membutuhkan waktu yang lama untuk menghitung skor antara kueri dan teks, meskipun hanya memeringkatkan (*reranking*) 1000 teks. Sementara itu, model BERT_{DOT} memerlukan memori yang lebih banyak karena representasi vektor padat dengan dimensi 768 harus disimpan terlebih dahulu. Model BERT_{CAT} hanya memerlukan memori yang sama dengan BM25.

Pada penelitian ini, interpretasi dari model tidaklah dibahas secara mendetail. paragraf berikut akan membahas sekilas mengenai interpretasi model BERT_{DOT} dan BERT_{CAT}.

Pada model BERT_{CAT}, interpretasi dapat dilakukan dengan *integrated gradients* (Sundararajan, Taly, & Yan, 2017). *Integrated Gradients* menghitung kontribusi dari setiap fitur (kata) pada hasil prediksi. Gambar 4.3 menunjukkan contoh interpretasi dari model BERT_{CAT} dengan *software* Captum.

Legend: ■ Negative □ Neutral ■ Positive				Word Importance
True Label	Predicted Label	Attribution Label	Attribution Score	
1	1 (0.99)	1	1.83	[CLS] kapan komando divisi ix banteng dibentuk ? [SEP] komando divisi ix banteng adalah suatu komando militer yang dibentuk pada masa perang kemerdekaan (1945 - 1950) di sumatera tengah yang wilayah operasinya meliputi empat provinsi sekarang , ya itu sumatera barat , riau , jambi dan kepulauan riau . komando divisi ix banteng mempunyai pasukan yang banyak karena adanya sekolah pendidikan opsi ##r di bukittinggi yang pendiriannya diprakarsai oleh para perwira didikan jepang dimasa pendudukan , bahkan salah satu pasukannya yaitu resimen 6 dianggap sebagai pasukan terbaik di sumatera . [SEP]
1	1 (0.99)	1	1.96	[CLS] apakah nama ibukota eti ##opia ? [SEP] kini eti ##opia merupakan negara berbentuk republik dan mengambil bagian secara aktif dal ##a , aktivitas - aktivitas kerja sama internasional . ibukota ##nya add ##is aba ##ba merupakan pusat administrasi kesatuan afrika (au) . [SEP]
1	1 (0.99)	1	1.11	[CLS] apakah yang dimaksud dengan revolusi ? [SEP] menurut kamus besar bahasa indonesia revolusi / re · vo · lu · si // revolusi / n * 1 perubahan ketatanegaraan (pemerintahan atau keadaan sosial) yg dilakukan dn ##g kekerasan (spt dn ##g perlawanan bersenjata) [SEP]
1	1 (1.00)	1	1.23	[CLS] apa akar dari semua kejahatan [SEP] secara pribadi , saya percaya akar dari semua kejahatan adalah ke ##ego ##isan dan ##ke ##ego ##isan diungkapkan dalam segala cara . sebaliknya adalah cinta yang mengorbankan diri sendiri . seseorang juga dapat berbicara tentang akar segala kejahatan dalam hal dusta . [SEP]

Gambar 4.3: Interpretasi dari model BERT_{CAT} dengan *integrated gradients*. Kata dengan warna hijau berarti kata tersebut berkontribusi positif terhadap hasil prediksi. Di lain sisi, kata yang berwarna merah berarti kata tersebut berkontribusi negatif terhadap hasil prediksi.

Model berasitekturkan BERT_{DOT} jauh lebih sulit untuk menginterpretasikannya. Hal yang dapat dilakukan adalah dengan menghitung nilai *importance* kata dengan menghitung hasil kali titik representasi vektor dari teks dengan representasi vektor masing-masing kata pada teks tersebut. Dengan hal ini, dapat nilai *importance* dari kata dapat diurutkan. Tabel 4.19 menunjukkan kueri, teks dan 5 kata penting dari kueri dan teks tersebut.

Tabel 4.19: Interpretasi dari model BERT_{DOT} dengan menghitung hasil kali titik antara vektor teks dengan vektor masing-masing kata pada teks tersebut. Hanya 5 kata dengan nilai *importance* tertinggi yang ditunjukkan.

Kueri	5 Kata Penting	Teks	5 Kata Penting
Kapan Petrus Lombardus lahir?	[lahir, petrus, kapan, lombardus, ?]	Petrus Lombardus mungkin dilahirkan di Novara; atau kemungkinan lainnya adalah di Lumello[7] (saat itu sebuah komune pedesaan, sekarang menjadi bagian dari Provinsi Novara, Piemonte), di barat laut Italia, dari suatu keluarga miskin.[8] Kelahirannya diperkirakan antara tahun 1095-1100.	[petrus, dilahirkan, keluarganya, lombardus, keluarga]
Dimana Jamie Richard Vardy lahir?	[lahir, richard, jamie, ?, dimana]	Jamie Richard Vardy (lahir dengan nama Gill; 11 January 1987) adalah pemain sepak bola Inggris yang bermain di klub Premiere League Leicester City dan tim nasional Inggris. Ia bermain sebagai striker, namun juga bisa bermain di sayap	[gill, lahir, richard, tim, sepak]

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

- Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate*.
- 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>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Diakses dari <http://arxiv.org/abs/1810.04805>
- Geiger, A., Antic, B., & He, H. (2022). *Lecture: Deep learning, university of tübingen*. <https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/autonomous-vision/lectures/deep-learning/>.
- Hofstätter, S., Althammer, S., Sertkan, M., & Hanbury, A. (2021). *Advanced information retrieval 2021 & 2022*. Diakses dari <https://github.com/sebastian-hofstaetter/teaching>
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... Yih, W.-t. (2020, November). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 6769–6781). Online: Association for Computational Linguistics. Diakses dari <https://www.aclweb.org/anthology/2020.emnlp-main.550> doi: 10.18653/v1/2020.emnlp-main.550
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian NLP. *CoRR*, *abs/2011.00677*. Diakses dari <https://arxiv.org/abs/2011.00677>
- 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>
- Lippe, P. (2022). *UvA Deep Learning Tutorials*. <https://uvadlc-notebooks.readthedocs.io/en/latest/>.

- Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.
Diakses dari probml.ai
- pi tau. (2023). An even more annotated transformer. *pi-tau.github.io*. Diakses dari <https://pi-tau.github.io/posts/transformer/> (Published on July 13, 2023)
- Potts, C., Ethayarajh, K., Karamcheti, S., Lee, M., Li, S., Li, X. L., ... Ògúnremí, T. (2023). *Cs224u: Natural language understanding*. <https://web.stanford.edu/class/cs224u/index.html>.
- Reimers, N., & Gurevych, I. (2020, 04). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*. Diakses dari <http://arxiv.org/abs/2004.09813>
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *CoRR*, *abs/1703.01365*. Diakses dari <http://arxiv.org/abs/1703.01365>
- Tsai, Y. H., Bai, S., Yamada, M., Morency, L., & Salakhutdinov, R. (2019). Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. *CoRR*, *abs/1908.11775*. Diakses dari <http://arxiv.org/abs/1908.11775>
- van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *CoRR*, *abs/1807.03748*. Diakses dari <http://arxiv.org/abs/1807.03748>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (p. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.
- Weng, L. (2018). Attention? attention! *lilianweng.github.io*. Diakses dari <https://lilianweng.github.io/posts/2018-06-24-attention/>
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press. (<https://D2L.ai>)
- Zhang, X., Ma, X., Shi, P., & Lin, J. (2021). Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv:2108.08787*.
- Zhang, X., Thakur, N., Ogundepo, O., Kamaloo, 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/tacl_a-00595 doi: 10

.1162/tac1_a_00595

LAMPIRAN

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.
 - Perbaiki 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.