

ABSTRAK

Franky Halim:

Skripsi

Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT

Pada era yang semakin modern, informasi menjadi bagian yang penting dalam kehidupan sehari-hari. Dalam mendapatkan informasi terdapat beberapa hal yang dapat dilakukan dimana salah satunya adalah dengan membaca. Dengan semakin banyaknya informasi yang tersedia di internet dapat membuat manusia kerepotan untuk terus mengikuti perkembangannya. Berita *online* juga merupakan salah satu sumber informasi yang ada di internet dengan jumlah yang sangat banyak serta topik yang beragam. Membaca keseluruhan informasi tersebut terkadang juga memerlukan waktu yang lama. Oleh karena itulah, diperlukan suatu pembuatan ringkasan dari informasi berita yang tersedia secara *online* untuk mengurangi waktu baca dan mendapatkan informasi yang relevan.

Pada penelitian ini akan dilakukan pembuatan ringkasan berita dengan memilih kalimat penting dari teks berita. Metode yang digunakan adalah *Bidirectional Encoder Representations from Transformers* dengan tambahan *transformer encoder layer*.

Berdasarkan hasil pengujian yang telah dilakukan, *pre-trained* model indolem/indobert-base-uncased dapat menghasilkan nilai *F1-Score* terbaik untuk ROUGE-1 sebesar 57.17, ROUGE-2 sebesar 51.27, dan ROUGE-L sebesar 55.20 pada referensi abstraktif serta ROUGE-1 sebesar 84.46, ROUGE-2 sebesar 83.21, dan ROUGE-L sebesar 83.40 pada referensi ekstraktif.

Kata kunci:

Text Summarization, Online News, Bidirectional Encoder Representations from Transformers.

ABSTRACT

Franky Halim:

Undergraduate Thesis

Automatic Extractive Summarization on Indonesian News using BERT Method

In this modern era, information has become an important part of everyday life. In getting information several things can be done where one of them is by reading. With the increasing amount of information available on the internet, it is difficult for humans to keep abreast of developments. Online news is also one of the sources of information on the internet with a very large number and various topics. Reading the whole information sometimes also takes a long time. Therefore, it is necessary to make a summary of the available online news to reduce reading time and obtain relevant information.

In this research, a summary of the news will be made by selecting important sentences from the news text. The method used in this research is Bidirectional Encoder Representations from Transformers with the addition of transformer encoder layer.

Based on the results of the tests that have been carried out, the pre-trained indolem/indobert-base-uncased model can produce the best *F1-Score* 57.17 for ROUGE-1, 51.27 for ROUGE-2, and 55.20 for ROUGE-L using abstractive reference and 84.46 for ROUGE-1, 83.21 for ROUGE-2, and 83.40 for ROUGE-L using extractive reference.

Keyword:

Text Summarization, Online News, Bidirectional Encoder Representations from Transformers.

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN.....	ii
LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS	iii
KATA PENGANTAR.....	iv
ABSTRAK	vi
ABSTRACT	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR	xi
DAFTAR PERSAMAAN	xii
DAFTAR TABEL.....	xiii
DAFTAR SEGMENT DATA.....	xiv
1. PENDAHULUAN	1
1.1 Latar Belakang Permasalahan.....	1
1.2 Perumusan Masalah	2
1.3 Tujuan Skripsi.....	3
1.4 Ruang Lingkup.....	3
1.5 Metodologi Penelitian	4
1.6 Sistematika Penulisan	4
2. LANDASAN TEORI	6
2.1 Tinjauan Pustaka	6
2.1.1 News Summarization.....	6
2.1.2 Indosum	6
2.1.3 Bidirectional Encoder Representations from Transformers (BERT)	7
2.1.4 Transformer	11
2.1.5 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)	13
2.2 Tinjauan Studi	14
2.2.1 Neural Summarization by Extracting Sentences and Words (Cheng & Lapata, 2016)	15
2.2.2 Peringkasan Ekstraktif Teks Bahasa Indonesia dengan Pendekatan Unsupervised Menggunakan Metode Clustering (Ismi & Ardianto, 2019)	15
2.2.3 Penerapan Recurrent Neural Network untuk Pembuatan Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia (Halim et al., 2020)	16

2.2.4 Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach (Nada et al., 2020)	17
3. ANALISIS DAN DESAIN SISTEM	18
3.1 Analisa Masalah	18
3.2 Analisa Data	18
3.2.1 Pengambilan Data	19
3.2.2 Pengolahan Data	19
3.3 Desain Sistem.....	19
3.3.1 BERT Embedding.....	20
3.3.2 Preprocessing	21
3.3.3 Training.....	22
3.3.4 Testing	23
3.3.5 Forward Pass BERT Summarization Model.....	25
3.3.6 Evaluasi ROUGE	26
3.4 Desain Program.....	26
4. IMPLEMENTASI SISTEM.....	27
4.1 Pengolahan dataset	28
4.2 Preprocessing.....	30
4.3 Konfigurasi google colaboratory	30
4.4 Inialisasi hyperparameter	31
4.5 Bert Data dan Dataloader	39
4.6 Training dan Testing.....	46
4.7 Transformer encoder	61
4.8 Evaluasi dengan ROUGE.....	63
4.9 Prediksi ringkasan	65
5. PENGUJIAN SISTEM	67
5.1 Pengujian Model	67
5.1.1 Pengujian Awal	69
5.1.2 Pengujian Kombinasi Learning Rate dan Dropout.....	75
5.1.3 Pengujian Token Type Ids.....	76
5.1.4 Pengujian Stacked Transformer Encoder	76
5.1.5 Pengujian Akhir.....	77
5.2 Pengujian program.....	80
6. KESIMPULAN DAN SARAN	84
6.1 Kesimpulan.....	84
6.2 Saran	84

DAFTAR REFERENSI	85
------------------------	----

DAFTAR GAMBAR

Gambar 1.1 Arsitektur model BERT dalam menghasilkan ringkasan	3
Gambar 2.1 Perbandingan struktur BERT	8
Gambar 2.2 Detail lengkap untuk struktur BERT Summarization	9
Gambar 2.3 Struktur transformer	11
Gambar 2.4 Diagram scaled dot product attention	12
Gambar 3.1 Alur sistem secara keseluruhan.....	20
Gambar 3.2 Alur persiapan data	21
Gambar 3.3 Alur preprocessing.....	21
Gambar 3.4 Alur training secara garis besar	23
Gambar 3.5 Alur testing secara garis besar	24
Gambar 3.6 Alur proses forward pass BERT Summarization Model.....	25
Gambar 3.7 Desain tampilan website	26
Gambar 5.1 Grafik loss konfigurasi awal IndoBERT	71
Gambar 5.2 Grafik loss konfigurasi awal Multilingual BERT	71
Gambar 5.3 Grafik F1-Score dari ROUGE-1 IndoBERT pada data training dan validation	72
Gambar 5.4 Grafik F1-Score dari ROUGE-1 Multilingual BERT pada data training dan validation	72
Gambar 5.5 Grafik F1-Score dari ROUGE-2 IndoBERT pada data training dan validation	73
Gambar 5.6 Grafik F1-Score dari ROUGE-2 Multilingual BERT pada data training dan validation	73
Gambar 5.7 Grafik F1-Score dari ROUGE-L IndoBERT pada data training dan validation.....	74
Gambar 5.8 Grafik F1-Score dari ROUGE-L Multilingual BERT pada data training dan validation	74
Gambar 5.9 Confusion Matrix Model Terbaik.....	79
Gambar 5.10 Contoh input file teks sebuah portal berita online	81
Gambar 5.11 Hasil ringkasan sistem yang menggunakan input file text	82
Gambar 5.12 Contoh input url sebuah portal berita online	83
Gambar 5.13 Hasil ringkasan sistem dari sebuah url portal berita online.....	83

DAFTAR PERSAMAAN

Persamaan 2.1 Position embeddings pada time step genap	11
Persamaan 2.2 Position embeddings pada time step ganjil	11
Persamaan 2.3 Scaled-dot product attention	12
Persamaan 2.4 ROUGE-N	14
Persamaan 2.5 Recall ROUGE-L.....	14
Persamaan 2.6 Precision ROUGE-L.....	14
Persamaan 2.7 F1-score ROUGE-L	14

DAFTAR TABEL

Tabel 4.1 Hubungan Segmen Program dan Desain Sistem	27
Tabel 4.2 Input features model BERT untuk proses training	46
Tabel 5.1 Pembagian data tiap fold	68
Tabel 5.2 Konfigurasi hyperparameter konstan.....	68
Tabel 5.3 Keterangan konfigurasi hyperparameter untuk pengujian model BERT.....	69
Tabel 5.4 Konfigurasi yang digunakan untuk pengujian awal model BERT.....	69
Tabel 5.5 Pengujian Awal pada Referensi Ekstraktif	70
Tabel 5.6 Pengujian Awal pada Referensi Abstraktif	70
Tabel 5.7 Pengujian Kombinasi Learning Rate dan Dropout pada Referensi Ekstraktif	75
Tabel 5.8 Pengujian Kombinasi Learning Rate dan Dropout pada Referensi Abstraktif.....	75
Tabel 5.9 Pengujian Token Type Ids pada Referensi Ekstraktif.....	76
Tabel 5.10 Pengujian Token Type Ids pada Referensi Abstraktif	76
Tabel 5.11 Pengujian Stacked Transformer Encoder pada Referensi Ekstraktif	77
Tabel 5.12 Pengujian Stacked Transformer Encoder pada Referensi Abstraktif	77
Tabel 5.13 Pengujian Akhir pada Referensi Ekstraktif	78
Tabel 5.14 Pengujian Akhir pada Referensi Abstraktif.....	78
Tabel 5.15 Perbandingan dengan metode-metode neural network yang sudah pernah diterapkan	80

DAFTAR SEGMENT DATA

Segmen Data 4.1 Contoh input features BERT untuk proses training	44
Segmen Data 5.1 Isi file teks berita yang diupload	82