



PERBANDINGAN ALGORITMA BART, RANDOM FOREST DAN HYBRID BART-RANDOM FOREST PADA *AUTOMATIC TEXT SUMMARIZATION*

MUHAMMAD ADIB ZAMZAM



PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Perbandingan Algoritma BART, Random Forest dan Hybrid BART-Random Forest pada *Automatic Text Summarization*” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Maret 2024

Muhammad Adib zamzam
G6501211006

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



RINGKASAN

MUHAMMAD ADIB ZAMZAM. Perbandingan Algoritma BART, Random Forest Dan Hybrid BART-Random Forest Pada Automatic Text Summarization. Dibimbing oleh AGUS BUONO dan TOTO HARYANTO

Data dan informasi berkembang secara kuantitatif dan kualitatif. Terdapat banyak teks pada internet dan pertumbuhan data menjadi lebih banyak dari yang dibutuhkan. Jumlah dokumen atau teks yang ada pada seluruh sumber sangat besar, maka pekerjaan merangkum menjadi sangat kompleks. *Natural Language Processing* (NLP) adalah subbidang pada computer science yang membahas pemrosesan dan analisis bahasa manusia. Pembahasan yang umum pada NLP yaitu pemrosesan percakapan, analisis bentuk kalimat, analisis sintaks, diskursus dan pembahasan terkait aplikasi teks seperti perangkuman (*summarization*), text generation, grammatical correction. *Automatic Text Summarization* (ATS) adalah salah satu tugas yang menantang pada NLP. ATS sangat sering digunakan pada text mining dan aplikasi analitis seperti *information retrieval*, *information extraction*, *question answering* dan sebagainya. ATS terdiri dari dua cara pendekatan umum yaitu abstraktif, ekstraktif dan *hybrid*. Pendekatan *hybrid* melakukan perangkuman dengan kombinasi dari abstraktif dan ekstraktif. Tujuan utama penelitian ini adalah menguji hasil performa algoritma BART dan Random Forest secara independen dan hasil performa secara kombinasinya dalam *automatic text summarization*. Digunakan algoritma Random Forest pada pendekatan ekstraktif, BART untuk pendekatan abstraktif dan kombinasi BART dan Random Forest untuk pendekatan hybrid. Penelitian menunjukkan bahwa secara individu, skor BART dan RF ROUGE cukup berbeda. Secara berturut-turut skor ROUGE RF pada R1, R2 dan RL adalah 1) 51.45 , 2) 45.52 dan 3) 54.58., skor ROUGE BART adalah 1) 32.78, 2) 16.17 dan 3) 32.19. Secara berturut-turut rata-rata pengukuran F ROUGE RF, BART dan RFxBART adalah 45.73, 21.38 dan 31.31. RF memiliki skor rata-rata tertinggi. ATS Hybrid RFxBART terbukti berkinerja lebih baik daripada BART default, tetapi lebih buruk daripada RF dalam hal skor ROUGE. Rata-rata ROUGE F RFxBART adalah 31,31. RFxBART memiliki skor sedang. Skor ini lebih baik daripada skor ROUGE default BART. RFxBART dapat menjadi alternatif pendekatan hybrid yang efektif.

Kata kunci: BART, NLP, perangkuman otomatis, random forest, rangkuman teks



SUMMARY

MUHAMMAD ADIB ZAMZAM. Text Summarization: BART, Random Forest, and Hybrid BART-RF Algorithm comparison. Supervised by AGUS BUONO and TOTO HARYANTO.

Data and information grows quantitatively and qualitatively. There is a lot of text on the internet and the growth of data is becoming more than needed. The number of documents or texts contained in all sources is very large, so the work of summarizing becomes very complex. Natural Language Processing (NLP) is a subfield of computer science that discusses the processing and analysis of human language. Common discussions in NLP includes conversation processing, sentence form analysis, syntax analysis, discourse and discussions related to text applications such as summarization, text generation, grammatical correction. Automatic Text Summarization (ATS) is one of the challenging tasks in NLP. ATS is often used in text mining and analytical applications such as information retrieval, information extraction, question answering and so on. ATS consists of two general approaches, which are abstractive, extractive and hybrid. The hybrid approach combines abstractive and extractive to do summarization. The main objective of this research is to test the performance results of the BART and Random Forest algorithms independently and their combined performance results in automatic text summarization. The Random Forest algorithm is used for the extractive approach, BART for the abstractive approach and a combination of BART and Random Forest for the hybrid approach. Research shows that individually, BART and RF ROUGE scores are quite different. Respectively the ROUGE RF scores on R1, R2 and RL are 1) 51.45, 2) 45.52 and 3) 54.58., the ROUGE BART scores are 1) 32.78, 2) 16.17 and 3) 32.19. Average measurements of F ROUGE RF, BART and RFXBART are 45.73, 21.38 and 31.31. RF had the highest average score. ATS Hybrid RFXBART was shown to perform better than default BART, but worse than RF in terms of ROUGE score. RFXBART's average ROUGE F is 31.31. RFXBART has a moderate score. This score is better than BART's default ROUGE score. RFXBART can be an effective alternative hybrid approach.

Keywords: BART, nlp, summarization, random forest, text summarization.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2024¹
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERBANDINGAN ALGORITMA BART, RANDOM FOREST DAN HYBRID BART-RANDOM FOREST PADA AUTOMATIC TEXT SUMMARIZATION

MUHAMMAD ADIB ZAMZAM

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Ilmu Komputer

**PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

IPB University

Judul Tesis : Perbandingan Algoritma BART, *Random Forest* Dan Hybrid
BART-*Random Forest* Pada *Automatic Text Summarization*
Nama : Muhammad Adib zamzam
NIM : G6501211006

Disetujui oleh

Pembimbing 1:
Prof. Dr. Ir. Agus Buono M.Si., M.Kom



Pembimbing 2:
Dr. Toto Haryanto S.Kom., M.Si.



Diketahui oleh

Ketua Program Studi:
Prof. Dr. Imas Sukaesih Sitanggang, S.Si., M.Kom.
NIP 19750130 199802 2 001



Dekan Fakultas Matematika dan Ilmu
Pengetahuan Alam:
Dr. Berry Juliandi, S.Si., M.Si.
NIP 19780723 200701 1 001



Tanggal Ujian:
14 Desember 2023

Tanggal Lulus:

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanaahu wa ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Oktober 2022 sampai bulan Juni 2023 ini ialah *Natural Language Processing*, dengan judul “Perbandingan Algoritma BART, Random Forest Dan Hybrid BART-Random Forest Pada Automatic Text Summarization”.

Terima kasih penulis ucapkan kepada para pembimbing, Prof. Dr. Ir. Agus Bueno M.Si., M.Kom dan Dr. Toto Haryanto S.Kom., M.Si. yang telah membimbing dan banyak memberi saran. Ucapan terima kasih juga disampaikan kepada moderator penguji, Dr. Medria Kusuma Dewi Hardhienata, S.Komp, dan penguji luar komisi pembimbing, Dr. Mushthofa S.Komp, M.Sc. Ungkapan terima kasih juga disampaikan kepada ibu, seluruh keluarga, serta seluruh teman kampus dan rekan karir atas segala dukungan dan doa sehingga penulis dapat menyelesaikan karya ilmiah ini.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Maret 2024

Muhammad Adib zamzam



DAFTAR ISI

DAFTAR TABEL	xi
DAFTAR GAMBAR	xi
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
1.5 Ruang Lingkup Penelitian	3
TINJAUAN PUSTAKA	4
2.1 <i>Automatic Text Summarization (ATS)</i>	4
2.2 Evaluasi Summarization	10
2.3 ROUGE	11
2.4 <i>Random Forest</i>	12
2.5 BART	15
III METODE	19
3.1 Tahapan penelitian	19
IV HASIL DAN PEMBAHASAN	24
4.1 Praproses dan Feature engineering	24
4.2 Perbandingan Model dan Hasil Rangkuman	25
4.3 Evaluasi dan Pembahasan	31
V SIMPULAN DAN SARAN	34
5.1 Simpulan	34
5.2 Saran	34
DAFTAR PUSTAKA	35
RIWAYAT HIDUP	40

Hak cipta milik IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR TABEL

1	Contoh hasil pra proses	20
2	Contoh fitur dataset pada model RF	21
3	Contoh hasil rangkuman BART (Lewis <i>et al.</i> 2020)	22
4	Perbandingan akurasi model RF dengan dan tanpa praproses	24
5	Cuplikan dataset baru	24
6	Tokenization pada satu titik data	25
7	Perbandingan waktu latih model RF dan BART	25
8	Contoh 1 Hasil rangkuman perbandingan 3 skenario	26
9	Rincian analisis Contoh 1	27
10	Contoh 2 Hasil rangkuman perbandingan 3 skenario	28
11	Rincian analisis Contoh 2	29
12	Contoh 3 Hasil rangkuman perbandingan 3 skenario	30
13	Rincian analisis Contoh 3	31
14	Perbandingan Skor ROUGE semua model	32
15	Perbandingan skor ROUGE RFxBART dengan model BART lain	33

DAFTAR GAMBAR

1	Ilustrasi dari proses transformasi teks El-Kassas <i>et al.</i> (2021)	4
2	Aplikasi yang mampu membantu aktivitas <i>preliminary research</i> dengan menggunakan kata kunci	6
3	Alur umum <i>extractive summarization</i> (El-Kassas <i>et al.</i> 2021)	7
4	Alur umum <i>abstractive summarization</i> (El-Kassas <i>et al.</i> 2021)	8
5	Alur umum <i>hybrid summarization</i> (El-Kassas <i>et al.</i> 2021)	9
6	Ilustrasi perhitungan ROUGE	12
7	Ilustrasi <i>Bagging</i> (Khan <i>et al.</i> 2019)	14
8	Arsitektur standar transformer (Vaswani <i>et al.</i> 2017) (a) dan perbandingan arsitektur GPT, BERT dan BART (Lewis <i>et al.</i> 2020) (b)	15
9	Ilustrasi dari proses transformasi teks (Lewis <i>et al.</i> 2020)	17
10	Tahapan penelitian pada sistem <i>summarization</i>	19
11	Cuplikan satu artikel dari dataset liputan 6	19
12	Contoh tahap <i>pre-training</i> pada BART (Lewis <i>et al.</i> 2020)	22
13	Ilustrasi <i>hybrid summarization</i>	23
14	Grafik ROC dan AUC model RF	26
15	Grafik perbandingan rata-rata skor F-measure ROUGE RF, BART dan RFxBART	32

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.